

A Thesis for the Degree of Ph.D. in Engineering

Non-Invasive Detection of Dementia Through Language
and Diagnosis Video Data: Machine Learning Methods
and Applications

September 2025

Graduate School of Science and Technology
Keio University

Chuheng Zheng

Contents

List of Tables	v
List of Figures	ix
Abstract	xiii
Acknowledgments	xvi
1 Introduction	1
1.1 Background	2
1.1.1 Definition of Dementia	2
1.1.2 Prevalence and Incidence	2
1.1.3 Society and Economy Burden	3
1.1.4 Clinical Classification of Dementia	4
1.2 Diagnosis of Dementia	6
1.2.1 Clinical Assessment	6
1.2.2 Imaging and Laboratory Tests	8
1.2.3 Limitations of Current Diagnostic Approaches	10
1.3 Machine Learning in Dementia Detection	11
1.3.1 Dementia Detection Using Neuroimaging Data	12
1.3.2 Dementia Detection Using Speech and Language Data	12
1.3.3 Dementia Detection Using Video Data	13
1.4 Outline and Contribution of the Dissertation	14
1.4.1 Challenges in Dementia Detection	14

1.4.2	Contribution of the Dissertation	17
2	An Evaluation on Information Composition in Dementia Detection Based on Speech	21
2.1	Introduction	22
2.2	Related Work	25
2.3	Dementia Detection Based on Language Models	36
2.3.1	Data Set	36
2.3.2	Pre-processing of Data Set	37
2.3.3	Baseline Methods	39
2.3.4	N-gram Language Model and PoS Tag Based Dementia Detection	40
2.3.5	Neural Network-Based Dementia Detection	42
2.4	Evaluation of Performance	45
2.4.1	Baseline Methods	45
2.4.2	PoS Sequences based Classification	46
2.4.3	AWD-LSTM Network	52
2.4.4	Attention-Based Network	53
2.4.5	Discussion	53
2.4.6	Error Analysis	55
2.5	Conclusions	58
3	Dementia Detection Based on Anomaly Detection from Text Embeddings using Autoencoders	59
3.1	Introduction	60
3.2	Related Work	62
3.3	Dementia Detection Based on Anomaly Detection	66
3.3.1	Language Model Embeddings	66
3.3.2	Autoencoders for Anomaly Detection	67
3.3.3	Dataset and Preprocessing	68
3.3.4	Anomaly Detection Approach	68

3.4	Experiment Settings and Results	70
3.4.1	Dementia Detection with Healthy Samples Only	70
3.4.2	Dementia Detection with Healthy and Dementia Samples	76
3.4.3	Autoencoder Variants	82
3.4.4	Error Analysis	90
3.5	Conclusion	92
4	Detecting Dementia from Face-related Features with Automated Computa-	
	tional Methods	93
4.1	Introduction	94
4.2	Related Work	95
4.3	Dataset and Facial Features	101
4.3.1	PROMPT Dataset	101
4.3.2	Raw Face Images	103
4.3.3	Facial Feature: the Face Mesh	105
4.3.4	Histogram of Oriented Gradients	106
4.3.5	Facial Feature: Action Unit	106
4.4	Experiment Designs and Settings	107
4.4.1	Cloze Test Experiments on Institutional Bias	107
4.4.2	Experiment Settings for 2D Facial Features	108
4.4.3	Experiment Settings for Face Mesh	110
4.4.4	Experiment Settings for HOG features	114
4.4.5	Experiment Settings for Action Units	115
4.5	Results and Evaluation	118
4.5.1	Cloze Test Experiments on Institutional Bias	118
4.5.2	Classification Using 2D Facial Features	122
4.5.3	Classification Using Face Mesh	124
4.5.4	Classification Using HOG Features	126
4.5.5	Classification Using Action Unit Features	127
4.5.6	Error Analysis	128

4.6	Conclusion and Future Directions	132
5	Conclusions and Future Work	133
5.1	Contributions	134
5.2	Conclusions	135
5.3	Applications and Future Work	140
6	Appendix	143
	Appendix A List of Author's Publications and Awards	145
A.1	Journals	145
A.2	Full Articles on International Conferences Proceedings	146
A.3	Articles on Domestic Conference Proceedings	146
A.4	Presentations at domestic meetings	147
A.5	Awards	148
	References	149

List of Tables

2.1	Comparison between our research and the referred research.	45
2.2	The performance of the experiments using CNN, LSTM, and CNN-LSTM models on the Pitt Corpus data set under 10-fold cross-validation. The results are reported as accuracy, precision, recall, and F1 score.	45
2.3	The performance of the experiments using CNN, LSTM, and CNN-LSTM models on the Pitt Corpus data set under Leave-One-Subject-Out (LOSO) cross-validation. The results are reported as accuracy, precision, recall, and F1 score.	46
2.4	Comparison of Physician and Patient Preferences in Diagnostic Tradeoff .	50
2.5	Classification Performance of the n -gram language models on the 2 Sets $\mathcal{P}$ and $\mathcal{T}$	54
2.6	Classification Performance of the AWD-LSTM Model on the 3 Sets $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{O}$	54
2.7	Classification Performance of the Attention-Based Model on the 3 Sets $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{O}$	55
2.8	Comparison of classification accuracy and interpretability approaches. . .	55
3.1	Average performance across latent dimensions: accuracy, ROC AUC, and class-specific precision, recall, and F1-score for the autoencoder trained on RoBERTa embeddings.	72
3.2	Average performance across latent dimensions: accuracy, ROC AUC, and class-specific precision, recall, and F1-score for the autoencoder trained on LLaMA 3.1 7B embeddings.	72

3.3	Average performance across latent dimensions: accuracy, ROC AUC, and class-specific precision, recall, and F1-score for the autoencoder trained on LLaMA 3.2 7B embeddings.	72
3.4	Cross-validation performance summary across latent dimensions when using RoBERTa embeddings, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	77
3.5	Cross-validation performance summary across latent dimensions when using LLaMA 3.1 embeddings, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	77
3.6	Cross-validation performance summary across latent dimensions when using LLaMA 3.2 embeddings, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	77
3.7	Performance summary across latent dimensions when using deep autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	83
3.8	Performance summary across latent dimensions when using denoising (noise level 0.1) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	83
3.9	Performance summary across latent dimensions when using denoising (noise level 0.2) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	84
3.10	Performance summary across latent dimensions when using denoising (noise level 0.3) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	84
3.11	Performance summary across latent dimensions when using denoising (noise level 0.4) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	84

3.12	Performance summary across latent dimensions when using variational autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.	85
3.13	Classification performance across latent dimensions using variational autoencoders trained on LLaMA3.1 embeddings.	86
3.14	Classification performance across latent dimensions using deep autoencoders trained on LLaMA3.1 embeddings.	86
3.15	Classification performance across latent dimensions using denoising autoencoders (noise level 0.1) trained on LLaMA3.1 embeddings.	87
3.16	Classification performance across latent dimensions using denoising autoencoders (noise level 0.2) trained on LLaMA3.1 embeddings.	87
3.17	Classification performance across latent dimensions using denoising autoencoders (noise level 0.3) trained on LLaMA3.1 embeddings.	87
3.18	Classification performance across latent dimensions using denoising autoencoders (noise level 0.4) trained on LLaMA3.1 embeddings.	88
3.19	Performance summary across latent dimensions using basic autoencoders with LLaMA3.1 embeddings.	88
3.20	Comparison of accuracy across different CI detection methods.	90
4.1	Demographic information of subjects by age and MMSE Score	102
4.2	Comparison of the different facial features: MediaPipe Face Mesh, HOG, Haar Cascades, and Facial Landmarks.	114
4.3	Cumulative explained variance for the top 1 – 16 principal components.	117
4.4	The accuracy performance when using different parts of the face to detect dementia.	122
4.5	Top 20 Features by Absolute Coefficient in the Classification using 2D Facial Features.	123
4.6	Dementia detection performance of deep learning models with different methods to add variations. Results are reported for accuracy, precision, and recall.	125

4.7	Classification performance in dementia detection using different numbers of frames of HOG features.	126
4.8	Classification performance of the action units under different settings. Segment length refers to the window size when segmenting the video samples.	127
4.9	Performance comparison of different methods.	128
5.1	Comparison of traditional tests and speech- and facial expression-based detection methods for dementia detection.	137
5.2	Summary of methods cited in Chapters 2, 3, and 4 with accuracy and evaluation remarks.	139

List of Figures

1.1	Landscape and context of the dissertation.	16
2.1	Examples of words with different lexical frequencies.	31
2.2	Overall pipeline of the research.	32
2.3	A flowchart of the pre-processing of the text data set. The original text is processed into 4 different instances that are used in the experiments. . . .	36
2.4	A flow chart of using two perplexities methods to detect dementia. . . .	40
2.5	The AWD-LSTM neural network architecture.	42
2.6	The transformer neural network architecture.	42
2.7	The perplexities difference between language models using only Part-of- Speech tags.	47
2.8	The perplexities difference between language models using Part-of-Speech tags and stop words.	47
2.9	Receiver Operating Characteristic curve of classification performance for using sequences of PoS tags and stop words.	50
2.10	The influence of stop words list size: from 20 most used words to 2000 most used words.	51
2.11	Difference in perplexity by MMSE score. Blue points represent correctly classified samples; red points are misclassified.	56
2.12	Misclassification rate by MMSE score. Patients with moderate MMSE scores have higher error rates.	56

3.1	Architecture of the autoencoder used for anomaly detection in dementia detection tasks. The encoder maps the input embeddings to a lower-dimensional latent space, while the decoder reconstructs the original embeddings from this latent representation.	67
3.2	Overview of the anomaly detection approach for dementia detection. The autoencoder is trained on healthy samples, and the reconstruction error is used as an anomaly detection signal to classify dementia samples.	69
3.3	Average accuracy across latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings.	73
3.4	Reconstruction loss distributions for healthy and dementia samples in fold 2, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.	74
3.5	Reconstruction loss distributions for healthy and dementia samples in fold 4, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.	75
3.6	Average accuracy across latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings, using two separate autoencoders for healthy and dementia samples.	78
3.7	Reconstruction loss distributions for healthy and dementia samples using two separate autoencoders, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.	79
3.8	Reconstruction loss distributions for healthy and dementia samples using two separate autoencoders, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.	80

3.9	Histogram of reconstruction losses for dementia and healthy samples using two separate autoencoders, with latent dimension 256. (a) shows the distribution for reconstruction errors under the autoencoder trained on dementia data, while (b) shows the distribution for reconstruction errors under the autoencoder trained on healthy data. The x-axis represents the reconstruction loss, and the y-axis represents the number of samples. . . .	81
3.10	Average accuracy across latent dimensions for different autoencoder variants, including baseline, deeper autoencoders, denoising autoencoders, and variational autoencoders.	85
3.11	Average accuracy across latent dimensions for different autoencoder variants trained on LLaMA3.1 embeddings, including baseline, deeper autoencoders, denoising autoencoders, and variational autoencoders. . . .	89
3.12	ROC curve for the best-performing autoencoder variant (denoising autoencoder with noise level 0.3) trained on LLaMA 3.1 embeddings. The area under the curve (AUC) is 0.8913, indicating good performance in distinguishing between healthy and dementia-related samples.	91
3.13	Difference in reconstruction error by MMSE score. Blue points represent correctly classified samples; red points are misclassified.	91
3.14	Misclassification rate by MMSE score. Patients with MMSE scores near the boundary ($MMSE = 27$) have higher error rates.	92
4.1	Camera and mic settings used in the data collection process.	101
4.2	An example of face mesh extracted using MediaPipe.	104
4.3	The cloze test experiments on the raw face images. The first row shows the original images, and the second row shows the masked images. . . .	109
4.4	An example of face mesh before and after normalization.	111
4.5	The overall pipeline of performing dementia detection using AUs. . . .	116
4.6	The difference of variance and mean of each action unit between healthy subjects and dementia patients.	117

4.7	An example of CAM heat map that shows the focusing area of the CNN model.	118
4.8	An example of zooming in image data processed by the tool provided in the PaddleHub.	119
4.9	An example of image data cropped using PaddleHub and OpenCV with its corresponding heat map. (a) The face image that is blurred to protect the privacy of the patient. (b) The heat map corresponding to (a) that shows which part contributes to the classification.	120
4.10	An example of face mesh extracted from the image that contains a face using Media Pipe.	120
4.11	Examples of heat maps of the same picture but pre-processed with different tools. (a) the face image that is blurred and pre-processed by PaddleHub. (b) the corresponding heat map of (a). (c) the face image that is blurred and pre-processed using face mesh. (d) the corresponding heat map of (c).	121
4.12	Error rate of the classification of dementia and healthy subjects using AUs. The x-axis shows the institution where the data was collected, and the y-axis shows the error rate of the classification.	129
4.13	Error rate of the classification of dementia and healthy subjects using face mesh. The x-axis shows the institution where the data was collected, and the y-axis shows the error rate of the classification.	130

Non-Invasive Detection of Dementia Through Language and Diagnosis Video Data: Machine Learning Methods and Applications

Chuheng Zheng
zheng@ohtsuki.ics.keio.ac.jp.
Keio University, 2025

Supervisor: Prof. Tomoaki Ohtsuki, Ph.D.
ohtsuki@ics.keio.ac.jp

Abstract

Alzheimer’s disease, as the most prevalent form of dementia, accounts for approximately 60–70% of all dementia cases. It is currently incurable with existing medical treatments. Timely diagnosis and intervention are proved to be the effective method to slow down the progression of dementia and improve the quality of life for patients and their families. Therefore, non-invasive and easily accessible methods for early detection are essential.

This dissertation explores the application of machine learning techniques to enhance early dementia detection, focusing on the analysis of language and diagnosis video data. Specifically, we investigated the use of various machine learning models, including long short-term memory networks, transformers, convolutional neural networks, and other models to identify dementia-related patterns in speech data from the DementiaBank dataset and video data from the PROMPT dataset. The goal is to develop non-invasive, easy-to-use and affordable methods that can assist in the diagnosis of dementia, thereby improving patient outcomes and facilitating timely interventions.

Chapter 1 introduces the background of dementia and its impact on individuals and society, emphasizing the need for early detection and the potential of machine learning to address this challenge. It outlines the main objectives of the dissertation, concentrating on developing non-invasive and easy-to-use methods for early dementia detection methods. It also provides an overview of the challenges we encountered and the contributions we

have made in overcoming these challenges, which mainly include data scarcity, lack of explanation, low data quality, and institutional bias.

Chapter 2 focuses on language-based dementia detection using language models adapted from natural language processing. Despite several studies demonstrating the potential of language models in dementia detection, it remains a challenge to provide the patients and clinicians with justifiable and validatable explanations. In this chapter, we propose a novel approach that separates the semantic and syntactic components of language, allowing for a more understandable analysis of dementia-related patterns. Distinctively, we proposed leveraging stop words and part-of-speech sequences, to capture the syntactic structure of language, while context words are used to capture the semantic content. By separately analyzing these two components, we aim to provide clearer post hoc explanations for the model’s predictions, providing validatable explanations for the patients and clinicians as to which part of the language is impaired. We evaluated several models, including traditional n-gram models, long short-term memory networks, and attention-based models. Specifically, three models were developed: one using only context words (semantic features), one using stop words and part-of-speech sequences (syntactic features), and one combining both types of features. Experimental results showed that both semantic and syntactic features contribute to dementia detection, while syntactic features contribute more significantly. The three models achieved accuracies of 70.00%, 76.16%, and 81.54%, respectively.

Chapter 3 leverages the capabilities of recent large language models to generate text embeddings and utilizes autoencoders for anomaly detection in dementia-related data. In this chapter, we proposed a zero-shot learning approach to detect dementia using healthy data only and a binary classification method using both healthy and dementia data. We explored the effectiveness of different text embeddings, including those from RoBERTa and LLaMA models, in capturing the nuances of dementia-related data. We experimented with various autoencoder architectures, including deep autoencoders, denoising autoencoders, and variational autoencoders, to assess their performance in detecting dementia. The results demonstrated that the choice of architecture and latent dimension significantly

impacts the performance of the autoencoder in detecting dementia. Denoising autoencoders, particularly with a noise level of 0.3, achieved the highest average accuracy, indicating that adding noise during training can help the model generalize better to unseen data.

Chapter 4 explores the potential of face-related features for dementia detection using the PROMPT dataset. In this chapter, we first identified and proposed methods to reduce the institutional bias in the PROMPT dataset, caused by limited data collection from several limited institutions. After identifying this bias, we extracted the three types of features from the diagnosis video recordings, including face mesh, histogram of oriented gradients features, and action units, and evaluated their effectiveness in distinguishing between dementia patients and healthy controls. Traditional machine learning and deep learning models were trained on these features to investigate their effectiveness. Among them, histogram of oriented gradients features achieved the highest accuracy of 79%, followed by action unit features with 71%, and face mesh features with 76%. These results suggest that facial information serves as a valuable indicator for automated dementia detection and highlight the importance of exploring non-verbal behavioral cues.

Chapter 5 summarizes the dissertation, highlighting its key points and contributions, and suggesting possible future work.

Acknowledgments

I would like to express my deepest gratitude to Professor Tomoaki Ohtsuki for his unwavering support, expert guidance, and valuable insights throughout the course of my research. His profound knowledge, thoughtful feedback, and encouragement played a critical role in the development and completion of this dissertation. I am particularly grateful for the opportunities he provided me to explore new ideas, engage with current research, and grow as an independent scholar.

I am also sincerely thankful to Associate Professor Mondher Bouazizi for his dedicated mentorship, technical advice, and continuous encouragement. His expertise in machine learning and data analysis greatly enriched the quality of this work. His attention to detail and willingness to discuss and refine various aspects of the research were essential to its successful completion.

I would like to extend my appreciation to all the members of the Ohtsuki Laboratory for their collaborative spirit, constructive discussions, and camaraderie. Their diverse perspectives and shared enthusiasm for our work created a stimulating and supportive environment that greatly enhanced my research experience.

Finally, I would like to thank my family, especially my parents, for their unwavering support and encouragement throughout this journey. Their belief in my abilities and constant motivation provided me with the strength to overcome challenges and stay focused on my goals.

Chuheng Zheng

Chapter 1

Introduction

1.1 Background

1.1.1 Definition of Dementia

Dementia is a clinical syndrome characterized by a progressive decline in cognitive functioning that interferes with an individual's ability to perform daily activities independently [1]. It involves deficits in multiple cognitive domains, including memory, language, executive function, attention, and visuospatial abilities [1, 2]. The condition is distinct from normal aging, which may involve mild changes in cognition that do not interfere with functional autonomy [3]. Moreover, dementia is differentiated from acute confusional states such as delirium, which typically have a rapid onset and are reversible with treatment of the underlying cause [4].

According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), dementia falls under the category of “Major Neurocognitive Disorders,” which are defined by significant cognitive decline from a previous level of performance in one or more cognitive domains, as evidenced by both subjective concern and objective assessment, and by the impairment's interference with independence in everyday activities [5]. The International Classification of Diseases, Eleventh Revision (ICD-11), similarly categorizes dementia under “Neurocognitive Disorders,” specifying a gradual and progressive deterioration of cognitive function [6].

1.1.2 Prevalence and Incidence

Dementia represents a major global health challenge due to its rising prevalence and the progressive, incurable nature of most of its forms. According to estimates from the World Health Organization (WHO), over 55 million individuals worldwide were living with dementia in 2021, with approximately 10 million new cases arising each year. Alzheimer's Disease International (ADI) similarly reports that dementia is among the leading causes of disability and dependency among older adults globally [7, 8].

Prevalence and incidence of dementia increase sharply with age. For individuals aged 65–69 years, global prevalence estimates range from 1% to 3%, while rates exceed 30% in

those aged 85 years and older. Women are disproportionately affected by dementia, both due to longer average life expectancy and possible sex-specific biological and hormonal risk factors [9]. Regional disparities are also significant. High-income countries generally report higher prevalence due to better diagnostic infrastructure and aging demographics; however, low- and middle-income countries (LMICs) are experiencing the most rapid increases in incidence as life expectancy rises and demographic transitions accelerate [10]. Socioeconomic factors play a key role in both risk and diagnosis. Lower educational attainment, limited access to healthcare, and fewer opportunities for cognitive and social engagement are all associated with increased dementia risk. Moreover, underdiagnosis and misclassification remain prevalent, particularly in resource-constrained settings, leading to underestimation of the true burden of the disease [11].

Demographic trends indicate that the global population is aging at an unprecedented rate. The proportion of individuals over 65 years of age is projected to rise from 10% in 2022 to 16% by 2050, with the fastest growth occurring in LMICs [12]. As a consequence, the number of people living with dementia is expected to increase substantially. Projections suggest that by 2030, the global prevalence of dementia will reach 57.4 million, and by 2050, it will surpass 152.8 million, nearly tripling from current levels [13]. This anticipated rise poses serious challenges for healthcare systems, particularly in regions where geriatric services, long-term care facilities, and trained healthcare personnel are insufficient to meet demand. Furthermore, these projections may be conservative, as they do not fully account for changes in risk factor profiles, such as increasing rates of diabetes, obesity, and cardiovascular disease, all of which are linked to elevated dementia risk [9, 11].

1.1.3 Society and Economy Burden

Dementia imposes a substantial economic burden on individuals, families, healthcare systems, and societies at large [14]. In 2019, the total global cost of dementia was estimated at USD 2.8 trillion, encompassing both direct medical expenditures and indirect costs

such as lost productivity and informal caregiving. This figure is projected to rise to USD 4.7 trillion by 2030 if effective interventions are not implemented [15].

Direct costs include expenditures related to medical care, diagnostic testing, pharmacological treatment, and institutionalization. These costs are especially high in high-income countries due to greater use of formal healthcare services. In contrast, in many LMICs, where institutional care is less accessible, the burden is disproportionately borne by families [16].

Informal caregiving accounts for approximately 50% of the total cost of dementia globally. Family members, often women, provide the majority of unpaid care, which includes assistance with activities of daily living, supervision, and emotional support. This caregiving role often leads to physical, emotional, and financial strain, particularly when combined with limited access to support services or respite care [16, 17].

At the macroeconomic level, dementia also results in lost workforce participation, reduced household income, and increased reliance on public healthcare and welfare systems. The cumulative effect contributes to a cycle of economic vulnerability that further complicates disease management and prevention. Given these challenges, dementia has been recognized as a public health priority by international health agencies, necessitating coordinated efforts in policy development, resource allocation, and research [18].

1.1.4 Clinical Classification of Dementia

Dementia is not a single disease entity but an umbrella term encompassing a range of disorders with diverse etiologies and neuropathologies. These disorders are typically classified based on their presumed cause and pathological features. The most common forms are as follows:

Alzheimer's Disease

Alzheimer's Disease (AD) is the most prevalent cause of dementia, accounting for approximately 60–80% of all cases. It is characterized pathologically by the accumulation of extracellular amyloid-beta plaques and intracellular neurofibrillary tangles composed

of hyperphosphorylated tau protein. Clinically, AD presents with insidious onset and gradual progression of memory impairment, followed by deficits in language, executive functioning, and orientation. Early-onset forms (before age 65) are often familial and associated with genetic mutations in APP, PSEN1, or PSEN2 genes, while late-onset forms are commonly linked to the APOE ϵ 4 allele [19,20].

Vascular Dementia

Vascular dementia results from cerebrovascular disease, including large-vessel strokes, small vessel disease, or chronic ischemic changes. The onset may be sudden or step-wise and is often accompanied by focal neurological signs. Cognitive deficits may vary depending on the vascular territory affected but commonly include executive dysfunction and slowed processing speed. Vascular pathology may also coexist with Alzheimer's pathology, leading to mixed dementia [21,22].

Dementia with Lewy Bodies

Dementia with Lewy bodies is characterized by the presence of alpha-synuclein aggregates in cortical neurons. Clinically, Dementia with Lewy bodies is marked by fluctuating cognition, visual hallucinations, REM sleep behavior disorder, and parkinsonism. Cognitive symptoms typically precede or occur concurrently with motor symptoms, distinguishing it from Parkinson's disease dementia [23,24].

Frontotemporal Dementia

Frontotemporal dementia refers to a group of disorders resulting from degeneration of the frontal and/or temporal lobes. It is more likely to present before the age of 65 and is characterized by early changes in behavior, personality, and/or language. Frontotemporal dementia is subdivided into behavioral variant frontotemporal dementia and primary progressive aphasia, which itself includes nonfluent/agrammatic and semantic variants. Unlike AD, memory is relatively preserved in early stages [25,26].

Mixed Dementia

Mixed dementia involves a combination of pathological features from more than one subtype, most commonly AD and vascular dementia. Mixed pathology is particularly prevalent in older individuals and may complicate clinical diagnosis and treatment [27].

Other Dementia Syndromes

Additional, less common forms of dementia include:

- **Parkinson's Disease Dementia** – develops in the context of long-standing Parkinsonism, typically after several years of motor symptoms [28].
- **Huntington's Disease Dementia** – arises from a genetic mutation in the HTT gene, leading to chorea, psychiatric disturbances, and cognitive decline [29].
- **Creutzfeldt-Jakob Disease** – a rapidly progressive prion disease characterized by spongiform changes in the brain, leading to severe cognitive and motor impairment [30].
- **Normal Pressure Hydrocephalus** – characterized by the clinical triad of dementia, gait disturbance, and urinary incontinence, often potentially reversible with cerebrospinal fluid shunting [31].

1.2 Diagnosis of Dementia

1.2.1 Clinical Assessment

Clinical assessment serves as the foundation of dementia diagnosis and typically represents the first step in the diagnostic pathway. It involves a structured and detailed evaluation of cognitive, behavioral, and functional changes over time, often incorporating both patient self-reports and informant accounts. The primary aim is to determine whether cognitive decline exists, whether it impairs daily functioning, and whether it is consistent with a neurodegenerative process or a potentially reversible condition.

Patient History and Informant Interview

A comprehensive clinical history is essential for identifying the onset, duration, and progression of cognitive symptoms. Key areas of inquiry include memory loss, language difficulties, disorientation, changes in executive function, and impaired judgment [32]. Behavioral and psychological symptoms, such as apathy, agitation, mood disturbances, hallucinations, or delusions, should also be documented, as they may provide diagnostic clues regarding the subtype of dementia (e.g., hallucinations are more common in dementia with Lewy bodies) [33].

Informant interviews are a critical component, as individuals with cognitive impairment often underestimate or lack insight into their deficits. Structured tools such as the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) [34] or the AD8 Dementia Screening Interview [35] are commonly used to obtain objective corroboration of cognitive and functional decline.

Cognitive Screening Instruments

Several brief, standardized cognitive screening tests are used to assess general cognitive functioning. These instruments are designed to identify individuals who require more comprehensive neuropsychological evaluation and to provide a preliminary index of cognitive impairment severity.

Mini-Mental State Examination (MMSE): One of the most widely used tools, the MMSE evaluates orientation, registration, attention, calculation, recall, language, and visuospatial skills. It has a maximum score of 30, with scores below 24 typically suggesting cognitive impairment. However, it is limited by ceiling effects, educational and cultural biases, and reduced sensitivity to executive dysfunction [36].

Montreal Cognitive Assessment (MoCA): The MoCA is more sensitive than the MMSE for detecting Mild Cognitive Impairment (MCI) and early dementia. It includes additional tasks evaluating executive function, abstraction, and delayed recall, which makes it particularly useful in early-stage disease or for non-Alzheimer dementias [37].

Alzheimer's Disease Assessment Scale – Cognitive Subscale (ADAS-Cog): Used primarily in clinical trials, the ADAS-Cog offers a more detailed evaluation of cognitive domains affected by AD and is sensitive to changes over time [38].

Clinical Dementia Rating (CDR) Scale: The CDR is used to stage dementia severity based on patient performance in six domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. It incorporates both patient and informant input and distinguishes between normal cognition (CDR = 0), MCI (CDR = 0.5), and increasing levels of dementia severity (CDR $\geq$ 1) [39].

Besides, neuropsychiatric symptoms are common in dementia and may precede, accompany, or complicate cognitive impairment. Evaluation of behavioral changes should include screening for depression, anxiety, apathy, irritability, delusions, and hallucinations. Tools such as the Neuropsychiatric Inventory (NPI) [40] and the Geriatric Depression Scale (GDS) [41] help quantify symptom severity and assess their impact on quality of life and caregiver burden. However, distinguishing between primary psychiatric conditions and early dementia is often clinically challenging, particularly in older adults, and requires careful longitudinal observation and integration of neuropsychological and neurological findings.

1.2.2 Imaging and Laboratory Tests

Neuroimaging and laboratory investigations play an essential role in the diagnostic evaluation of dementia. These tools are used to identify structural, functional, or molecular abnormalities associated with neurodegeneration and to exclude potentially reversible causes of cognitive impairment. They also contribute to the differentiation among dementia subtypes and the confirmation of specific pathophysiological processes, particularly in research and specialized clinical settings [42].

Structural neuroimaging is recommended in all patients undergoing evaluation for dementia. The primary modalities are Magnetic Resonance Imaging (MRI) and Computed Tomography (CT).

MRI is the preferred modality due to its superior spatial resolution and sensitivity to soft tissue contrast. It provides detailed information on patterns of brain atrophy, vascular lesions, and other structural abnormalities. In AD, characteristic findings include medial temporal lobe atrophy, especially of the hippocampus and entorhinal cortex. Visual rating scales, such as the Scheltens scale for hippocampal atrophy, are commonly used in both clinical and research contexts [43].

CT, while less sensitive than MRI, remains useful in clinical settings where MRI is contraindicated or unavailable [44]. It is effective for detecting gross structural changes such as large-vessel infarcts, mass lesions, hydrocephalus, or significant atrophy. Both CT and MRI are critical for excluding secondary causes of dementia such as tumors, subdural hematomas, or normal pressure hydrocephalus.

Positron Emission Tomography (PET) is increasingly utilized in the evaluation of dementia, particularly for its ability to visualize metabolic and molecular processes in the brain. PET imaging can assess glucose metabolism using fluorodeoxyglucose (FDG-PET) or visualize amyloid-beta and tau protein deposition using specific radiotracers [45, 46]. FDG-PET is particularly useful in differentiating AD from other dementias, as it typically shows reduced glucose metabolism in the parietal, temporal, and posterior cingulate regions in AD. Amyloid PET imaging can confirm the presence of amyloid plaques, which is a key diagnostic criterion for AD.

In recent years, blood-based biomarkers have gained considerable attention as a promising alternative to cerebrospinal fluid (CSF) analysis and advanced neuroimaging for the diagnosis of AD and related dementias. The development of reliable plasma biomarkers offers the potential for non-invasive, cost-effective, and scalable screening tools that could be implemented in both clinical and population-level settings [47]. This is especially important in primary care and resource-limited contexts, where access to CSF analysis or PET is often restricted.

1.2.3 Limitations of Current Diagnostic Approaches

AD, the most prevalent form of dementia, remains a progressive and incurable neurodegenerative disorder. Once the pathological processes are underway, characterized by the accumulation of amyloid-beta plaques and tau neurofibrillary tangles, clinical progression typically follows a trajectory of gradual but irreversible cognitive and functional decline. To date, no available treatment has demonstrated the capacity to halt or reverse the neurodegenerative cascade once it reaches a symptomatic stage [48]. As such, a diagnosis made at a moderate or late stage, while helpful for care planning and symptom management, often provides limited clinical benefit in altering the disease course.

Existing diagnostic modalities, although increasingly accurate, present substantial limitations in terms of feasibility, accessibility, and patient acceptability. Neuroimaging techniques such as MRI and PET are resource-intensive, time-consuming, and often limited to specialized centers. Similarly, CSF analysis, while informative, involves invasive lumbar puncture procedures that may be contraindicated or refused by patients due to discomfort or perceived risk [49]. These diagnostic approaches may also raise concerns regarding data protection and patient privacy, especially when integrated into centralized electronic health records or subjected to secondary research use. The high cost and limited scalability of these tools further restrict their application in large-scale population screening or routine primary care settings.

Given these constraints, there is increasing consensus that the most effective strategy for managing dementia, particularly AD, lies in early detection and timely intervention. Pathophysiological changes in the brain may begin decades before the onset of overt clinical symptoms [50]. Identifying individuals in the preclinical or prodromal phases of the disease, such as during MCI, offers a critical window of opportunity to implement therapeutic, lifestyle, or psychosocial interventions that may delay progression, optimize quality of life, and reduce the burden on caregivers and healthcare systems.

Early detection also facilitates more informed decision-making regarding future care, legal and financial planning, and participation in clinical trials aimed at disease modification. Therefore, there is a growing need for non-invasive, low-cost, and privacy-preserving diagnostic methods that can be integrated into routine screening protocols, ideally at the community or primary care level. The development and validation of such approaches represent a critical priority in contemporary dementia research and public health policy.

1.3 Machine Learning in Dementia Detection

Machine learning has been achieving great success in many fields, such as computer vision, Natural Language Processing (NLP), speech processing, generative models, etc [51]. In computer vision, deep learning models, particularly Convolutional Neural Networks (CNNs), have revolutionized image classification, object detection, and segmentation tasks, achieving state-of-the-art performance on benchmarks such as ImageNet [52]. Similarly, in NLP, transformer-based architectures like BERT and GPT have set new standards for tasks such as text classification, translation, and question answering [53, 54].

In recent years, machine learning and deep learning have emerged as a powerful tool for analyzing complex medical data. The ability of deep learning models to automatically learn and extract relevant features from raw data without the need for manual feature engineering makes them highly suitable for medical image analysis, including the detection of neurological disorders such as dementia. Moreover, the advent of deep learning has been facilitated by the availability of large datasets, advances in computing power, and improved algorithmic frameworks, making it feasible to apply these methods in clinical settings [55].

1.3.1 Dementia Detection Using Neuroimaging Data

Deep learning-based dementia detection involves training neural networks on large datasets of brain images to recognize patterns and features indicative of dementia. These models can be trained using various types of neuroimaging data, including structural MRI, functional MRI (fMRI), and PET scans, as well as non-imaging data such as clinical records and genetic information. Structural MRI provides high-resolution images of brain anatomy, allowing for the identification of atrophy patterns and other structural changes associated with dementia. Functional MRI captures brain activity by measuring changes in blood flow, offering insights into functional connectivity and network disruptions that occur in dementia. PET imaging, on the other hand, allows for the detection of amyloid-beta plaques and tau protein tangles, hallmark pathologies of AD [56–58].

Despite these advances, challenges remain in applying machine learning to dementia detection, including limited availability of labeled imaging data and the reluctance of patients to undergo certain procedures. Imaging is often expensive and time-consuming, leading to smaller datasets compared to other domains such as image classification or NLP. Additionally, the invasive nature of some imaging techniques, such as PET and CSF analysis, limits their widespread use in routine clinical practice and may cause discomfort and reluctance among patients.

1.3.2 Dementia Detection Using Speech and Language Data

Speech and language data offer a promising avenue for dementia detection, as changes in communication patterns can be early indicators of cognitive decline. Machine learning techniques can be applied to analyze various aspects of speech and language, including lexical choice, syntactic structure, and prosody, to identify subtle markers of dementia. For instance, studies have shown that individuals with dementia may exhibit reduced vocabulary diversity, increased use of filler words, and changes in speech fluency [59]. By leveraging these linguistic features, machine learning models can be trained to classify speech samples as indicative of dementia or healthy aging.

Using speech and language data for dementia detection has several advantages. First, it is a non-invasive and cost-effective method that can be easily integrated into routine clinical assessments. Second, speech data can be collected in naturalistic settings, such as during clinical interviews or phone conversations, making it more accessible for patients who may have mobility or transportation challenges. Third, speech analysis can provide real-time feedback on cognitive changes, allowing for timely intervention and monitoring of disease progression.

1.3.3 Dementia Detection Using Video Data

Video data, particularly from clinical interviews, provide a rich source of information for dementia detection. Machine learning techniques can analyze visual features such as facial expressions, eye gaze, and body language to identify behavioral and emotional changes associated with dementia [60]. For example, individuals with dementia may exhibit reduced facial expressiveness, impaired eye contact, and altered body posture. By combining visual features with speech and language analysis, machine learning models can achieve more robust and accurate classification of dementia.

Using video data for dementia detection has several advantages. First, it captures both verbal and non-verbal cues, providing a more comprehensive assessment of cognitive and emotional functioning. Second, video data can be collected in naturalistic settings, allowing for more authentic interactions between patients and clinicians. Third, the integration of visual features with speech and language analysis can enhance the accuracy of dementia classification, as it captures a broader range of behavioral changes that may not be evident in speech alone.

The dementia detection methods based on speech and video data have great potential to improve early diagnosis and monitoring of dementia. By leveraging machine learning techniques, these methods can provide objective and quantifiable measures of cognitive decline, enabling clinicians to make more informed decisions about patient care. Most importantly, these methods are non-invasive, cost-effective, and easy-to-use, reducing the reluctance of patients to undergo diagnostic procedures. As a result, they can be integrated

into routine clinical practice and encourage early intervention and support for individuals with dementia.

1.4 Outline and Contribution of the Dissertation

1.4.1 Challenges in Dementia Detection

The goal of this dissertation is to develop non-invasive, easy-to-use and affordable methods for massive dementia screening. To achieve this goal, we focus on addressing four key challenges in dementia detection: data scarcity, lack of explanations, low data quality, and institutional bias.

First, data scarcity is a significant challenge, particularly in the medical domain, where obtaining large, labeled datasets is often difficult due to privacy concerns, ethical considerations, social stigma, and the high cost of data collection. This scarcity can lead to overfitting and poor generalization of machine learning models, limiting their applicability in real-world clinical settings. Patients with dementia can lose autonomy and may not be able to provide informed consent for data collection, leading to ethical dilemmas in obtaining representative datasets.

Second, the lack of justifiable and validatable explanations from the machine learning models poses a challenge in clinical applications. Many deep learning models, such as neural networks, are often considered “black boxes”, making it difficult for clinicians to understand how the model arrives at its predictions. Despite many researches have been conducted to improve the interpretability of machine learning models, such as using attention mechanisms or feature importance scores, these explanations may not align with the needs of clinicians and patients [61]. This lack of transparency can hinder trust and acceptance among healthcare professionals and patients, as they may be reluctant to rely on models that do not provide validatable explanations for their decisions.

Third, low data quality and institutional bias can affect the performance of machine learning models. Data collected from different institutions may vary in quality, leading

to inconsistencies and institutional bias in model training and evaluation. These inconsistencies and bias can make the models to perform perfectly on the dataset, however, this high performance does not generalize to other datasets or real-world settings, because the models are learning from the bias instead of the dementia biomarkers. Several researches have reported institutional bias in dementia datasets, such as [62, 63].

In this dissertation, our main aim is not to achieve higher accuracy in dementia detection. This is because medical datasets are usually private, small, imbalanced, and have institutional bias. Trying to achieve high accuracy in such a context may lead to overfitting on specific datasets and poor generalization to real-world applications. Many works have been reporting the institutional bias in dementia detection. [62] exposed underlying bias in Pitt corpus acoustic data, reporting very high accuracy using silenced segments of the audio. [63] explicitly reported their results were affected by institutional bias, even though the accuracy is 91%. Therefore, instead of pursuing higher accuracy, we aim at addressing the challenges of data scarcity, lack of explanations, low data quality, and institutional bias. We report accuracy as a relative metric to show if our methods work in addressing these challenges.

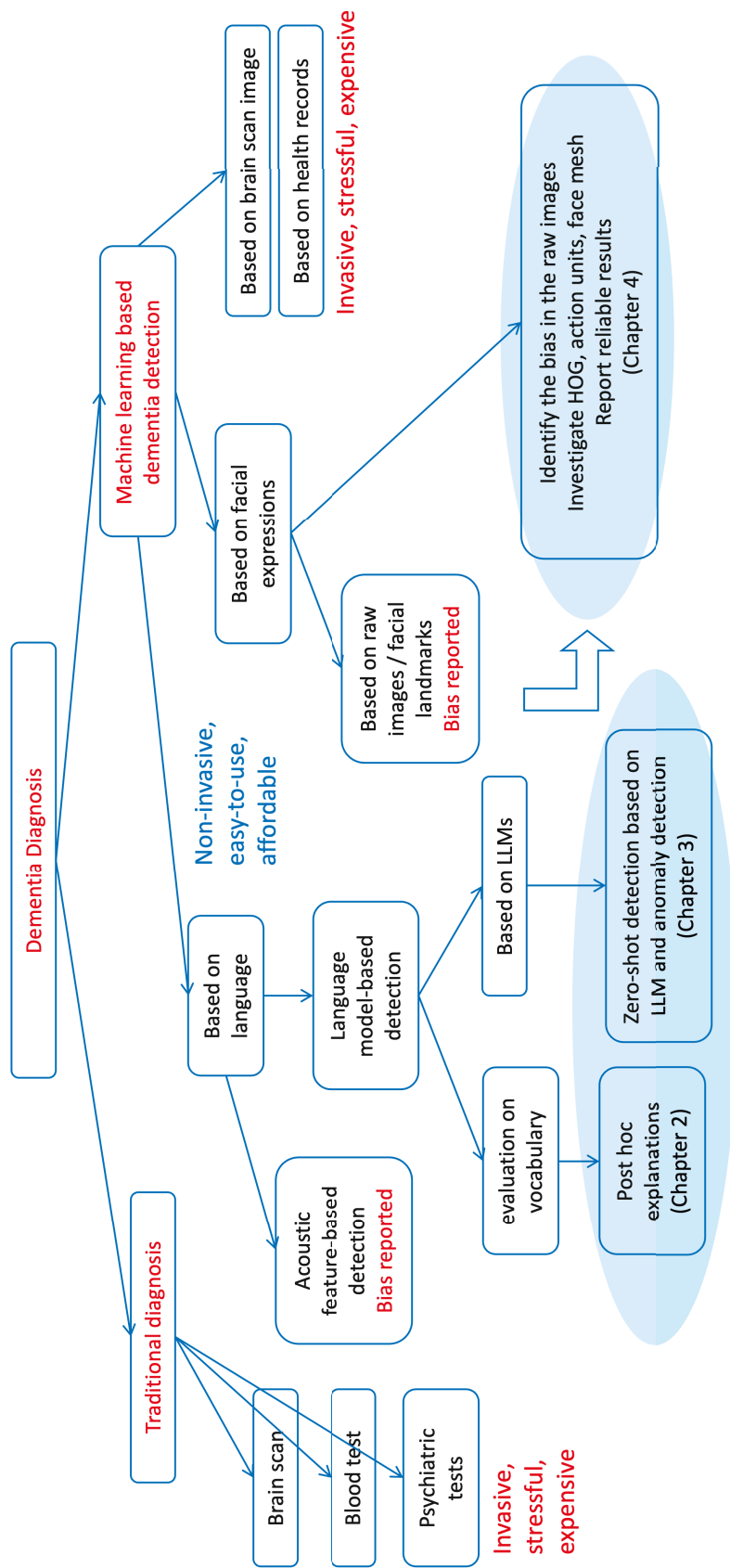


Figure 1.1. Landscape and context of the dissertation.

1.4.2 Contribution of the Dissertation

The dissertation consists of 5 chapters. Chapters 2, 3, and 4 proposed and investigated different approaches for early dementia detection, aiming to overcome the challenges explained in Sec. 1.4.1, while Chapter 5 summarizes the key findings and contributions of the research. Chapters 2 and 3 focus on language data, aiming to alleviate data scarcity problems and post hoc explanation of dementia detection models. Chapter 4 focus on diagnosis video data collected in the diagnosis interviews, aiming to identify and address the challenges of low data quality and institutional bias in dementia detection. Chapter 5 concludes the dissertation by summarizing the key findings and contributions of the research, discussing the implications for early dementia detection, and outlining potential avenues for future work in this critical area of healthcare.

Specifically, this dissertation makes several key contributions to the field of dementia detection. Chapter 2 contributes to the field of dementia detection by improving post hoc explanations and evaluations on the language model and perplexity based methods. Specifically, we proposed a novel method to separate the language components into syntactic and semantic components, and we systematically evaluate the contribution of each component to the classification performance. We show that the syntactic component of the language is a strong indicator of dementia, and we demonstrate that explicitly embedding the syntactic component into the text can improve classification accuracy. Our findings suggest that both syntax and semantics are important for detecting dementia. The novelty of our work lies in the separation of syntactic and semantic components in language data, which allows for a more detailed analysis of the individual composition of languages. By making such an analysis, we can know: 1. Which component of the language is contributing to the task of dementia detection. 2. Which aspects, either the vocabulary-related functions or the grammar-related functions, or both of the patients are impaired. This enables us to identify specific language deficits in patients, providing justifiable and validatable explanations for the diagnosis that can be understood, trusted, and validated by patients and clinicians.

Chapter 3 contributes to the field of dementia detection by proposing a novel approach that leverages autoencoders to perform zero-shot dementia detection using healthy language data only and binary classification using healthy and dementia language data. We systematically evaluate and visualize the performance of various autoencoder architectures and latent dimensions. The results indicate that the choice of embedding type has a more significant impact on classification performance than the latent dimension size or architecture complexity. Our findings suggest that autoencoders can be a valuable tool for dementia detection, particularly when using advanced embeddings like LLaMA. The novelty of our work lies in the use of autoencoders to reconstruct the text embeddings from large language models for zero-shot detection of dementia, which to our knowledge has not been explored before in dementia detection area. This approach, by employing an anomaly detection framework, allows for zero-shot detection using healthy language data only, which reduces the need for large, labeled dementia datasets. Besides, the use of text embeddings from large language models also has the potential to provide more understandable and validatable explanations for the diagnosis, which can be trusted and validated by patients and clinicians.

Chapter 4 contributes to the field of dementia detection by investigating the potential of using diagnosis video data. In this chapter, we first used explainable AI methods to analyze the video data and identify the potential institutional bias in the data by conducting a cloze test experiment. To address this bias, we proposed several methods and pipelines to normalize and process the video data. Additionally, after addressing the bias, we explored the potential of using facial features such as face mesh, HOG features, and AUs in dementia detection. Specifically, for the action unit data, we proposed a novel method that combines segmentation and voting to improve the detection accuracy. We adapted and improved the method in [64] through segmentation and voting. We also validated this method on a relatively large dataset, i.e., PROMPT dataset. For the face mesh data, we manually manipulate the 3-dimensional mesh to address the challenge of inconsistent postures and orientations in the video data. We investigated the effect of both normalization and regularization on the classification performance. The results show that the facial

features have the potential to work as an indicator in dementia detection, while protecting the privacy of the patients at the same time. Our approach addresses the challenges of low data quality and institutional bias by employing explainable AI methods to identify and mitigate bias, and by leveraging facial features that can be extracted from video data without requiring invasive procedures. The novelty of our work lies in the identification and mitigation of institutional bias in video data, which is common in medical datasets, as well as the exploration of facial features such as face mesh and action units for dementia detection. By addressing the bias and reporting reliable performance of facial features, we contribute to the understanding of how video data can be effectively utilized in dementia detection, while also protecting patient privacy.

Fig. 1.1 illustrates the landscape and context of the dissertation, highlighting the key contributions and methodologies employed in each chapter. In this study, we specifically used text-form language data and facial features including face mesh and action units for the following reasons:

- **Text-form language data:** We focus on text-form language data because it is informative, free from institutional bias, and can provide rich insights and explanations on the symptom-level, which can guide therapeutic interventions. Dementia is famous for its impact on language abilities, and analyzing language data allows us to capture those changes effectively. Text-form language data is especially favored because audio-form language data is often affected by the recording environment, such as background noise, which can introduce bias and variability in the data [62]. Besides, text-form language data has better potential to provide explanations for the diagnosis as to what specific language deficits the patients have, which can be understood, trusted, and validated by both patients and clinicians, especially with the recent advancements in LLMs.
- **Facial features:** We focus on facial features, including face mesh and action units in the context of video data, because these features can capture expression and movement patterns that are indicative of cognitive decline, while reducing the dimensionality of the data and avoiding institutional bias problem. Face mesh and

action units are rich in information and can provide insights into the emotional and behavioral changes embedded in the video data. They are especially favored over raw video data because raw video data is often affected by the recording environment, such as background, inconsistent postures and orientations, lighting conditions, and other factors that can introduce bias and variability in the data. Besides, lack of sufficient amount of video data is also hindering the use of raw video data in dementia detection. Raw video data has high dimensionality, this can cause overfitting when samples are limited, and it is also computationally expensive to process. By focusing on facial features, we can reduce the dimensionality of the data while still capturing the essential information needed for dementia detection.

Chapter 2

An Evaluation on Information

Composition in Dementia Detection

Based on Speech

2.1 Introduction

Research on the detection of dementia has attracted substantial attention in recent years. This is because establishing effective prevention strategies for AD requires detecting its pathology several years before the appearance of clinical symptoms [65]. From a medical standpoint, dementia is not a single disease but a broad term that encompasses a range of medical conditions. AD alone accounts for over 60% of all dementia cases [66,67]. While some types of dementia may be reversible if detected early, the majority are incurable. Nevertheless, expert assessment and early diagnosis can slow the disease's progression. Furthermore, early detection can help caregivers and family members better understand previously confusing behaviors exhibited by the patient. Due to the increasing importance of timely diagnosis, scientific interest in developing innovative methods for detecting dementia continues to grow. Accordingly, numerous studies have been conducted in this field [68–70].

Dementia testing can take many forms, ranging from cognitive evaluations and laboratory tests to brain scans and imaging techniques [71, 72]. Traditional methods for diagnosing dementia rely heavily on clinical assessments, including neuropsychological testing, patient history, and physical examinations. Imaging techniques such as MRI, PET, and CT have also been employed to detect structural and functional brain changes associated with dementia. Blood test, such as the analysis of amyloid-beta and tau proteins, have also been explored as potential biomarkers for AD, the most common form of dementia. However, these approaches require specialized equipment, can be expensive and are not always accessible in resource-limited settings. Besides, these methods, while effective to some extent, are often invasive, time-consuming, and stressful.

Therefore, automating diagnostic procedures can reduce costs and improve accessibility for the general population. Consequently, the scientific community has explored various strategies to automate dementia diagnosis. In particular, recent advances in Artificial Intelligence (AI) offer promising opportunities. AI has demonstrated the capacity

to detect subtle patterns in diverse types of data [52, 73, 74] and has the potential to significantly lower diagnostic costs. As a result, there is a growing interest in developing automated, reliable, and scalable methods for early dementia detection.

Among various types of data, speech is considered a valuable source of clinical information. Human speech is closely tied to cognitive status and is frequently used as a fundamental resource in mental health assessments. Speech data are also non-invasive, natural to collect, and widely available. Language patterns are indicative of cognitive health and can reflect the decline in cognitive function. Therefore, speech can be leveraged in the development of assistive technologies [75]. Dementia often leads to language impairments, such as difficulties in word-finding, comprehension, speech accuracy, and coherence [59]. Additionally, language processing relies on other cognitive functions, especially executive functioning, to ensure that communication is coherent and meaningful. Executive functions also play key roles in decision-making, strategic planning, and problem-solving, which are essential components of effective communication [76].

In recent years, researchers around the globe have been increasingly employing NLP and machine learning techniques to detect dementia based on speech and language data [76]. Language serves as a promising marker for early dementia detection. However, linguistic analysis is inherently complex, time-consuming, and typically demands manual effort by trained professionals. Advances in machine learning based speech and language processing technologies bring three key advantages. First, they enable the development of reliable tools for distinguishing between speech samples from individuals with dementia and those without. Second, they make it possible to quantify the severity and stages of dementia. Third, these technologies are non-invasive, cost-effective and easy-to-use [76–78]. Automated speech analysis systems are generally more acceptable to patients when facilitated by a human-like presence, such as a real person or an interactive robot. In contrast, ambient sensors or passive devices, which lack meaningful interaction, are less well-received [79]. As a result, current systems typically focus on analyzing natural, unscripted communication rather than fixed or scripted text.

Regarding the domain of dementia detection, a few publicly available datasets have enabled researchers to experiment with different models. These include DementiaBank¹ and Dem@Care², which provide data in various formats, such as audio, video, and transcribed text from dementia patients and control groups. Among these, speech data, either in audio or transcribed text format, are especially informative and have attracted the most research attention. Many studies have focused on processing such data, whether as audio signals or as text, for dementia detection [78,80,81], supported by advancements in NLP and audio signal processing. A variety of techniques for text classification have been proposed in the literature [78,82]. Whether the task involves sentiment analysis, hate speech detection, or bot identification [83], the general methodology remains similar: extract features from the text and apply machine learning or deep learning models to identify the target class. Applying these techniques to dementia detection has shown encouraging results [84–86].

However, we argue that there are important distinctions between tasks like opinion mining and the diagnosis of dementia when employing text classification techniques. In tasks such as sentiment analysis or hate speech detection, most relevant information is found directly in the content and meaning of words. This is not always the case in dementia detection, where linguistic structure, coherence, hesitation patterns, and other non-semantic features often carry more diagnostic value. Therefore, the future of dementia detection lies in the integration of deep learning with other emerging technologies, such as wearable devices and mobile health applications, to enable continuous monitoring and early detection of cognitive decline. By combining deep learning algorithms with real-time data from wearable sensors, it may be possible to develop personalized monitoring systems that can detect early signs of dementia before clinical symptoms become apparent. This approach could facilitate proactive interventions and lifestyle modifications to mitigate the progression of dementia.

¹<https://dementia.talkbank.org/>

²<https://demcare.eu/datasets/>

2.2 Related Work

Due to the factors that cause dementia and the symptoms shown in the patients are multimodal, the past years have seen different approaches being proposed for early dementia detection.

Nguyen et al. [87] explored the use of ChatGPT to assist in diagnosing Alzheimer’s Disease (AD) using electronic health records that include MRI data and cognitive test scores. They tested both zero-shot and multi-shot prompting strategies and find that multi-shot prompting, especially when combining both data types (MRI scan and cognitive tests), gave the best results. Their best model achieves 94.6% accuracy and shows good confidence calibration. This study shows that large language models can be useful tools for supporting clinical decision-making in AD diagnosis. However, this study is still invasive as it relies on MRI scans and cognitive tests, which may not be practical for the purpose of non-invasive dementia detection.

Shahin et al. [88] propose a zero-shot speech-based Cognitive Impairment (CI) detection method using the Qwen2-Audio AudioLLM, a model capable of processing both audio and text inputs. Their approach utilizes prompt-based instructions to classify speech samples as indicative of normal cognition or CI, eliminating the need for manual annotation and training on specific datasets. However, the model’s performance is not satisfying with an unweighted average recall of 57.5%.

Farzana et al. [89] presents a comprehensive evaluation of Domain Adaptation (DA) techniques for improving dementia detection from spoken language across heterogeneous datasets. Recognizing the limitations posed by small and varied health-related speech corpora, the authors investigate feature-based and instance-based DA methods to enhance model generalizability across both task-oriented and conversational domains. Using multiple datasets, they demonstrate that feature-augmented DA approaches, especially MULTIAUGMENT, consistently outperform baseline models and show up to 22% improvement in cross-domain accuracy.

A study by Roark [90] annotated a few speech features and aligned them by using NLP and automatic speech recognition tools. The same features are also extracted manually by human annotation. They studied the differences in the performance between the automated feature extraction and its human-annotated version. They evaluated 74 speech recordings to classify between MCI and healthy subjects. Their model of the best performance reached an Area Under the Curve (AUC) of 0.86 by combining speech and language features and cognitive test scores.

Zhu et al. [91] explored different transfer learning techniques for dementia detection, which involved using pre-trained models and fine-tuning them on the dementia data set. They conducted experiments that utilized wav2vec and BERT [92] to analyze both semantic and non-semantic information from patients' speech data for dementia detection. They developed a basic WavBERT model for semantic analysis and extended versions that preserved non-semantic features by incorporating inter-word pauses and fine-tuning through a conversion network.

Jarrold et al. [93] used a data set that included semi-structured interviews from 9 healthy individuals, 9 individuals with AD, 9 individuals with frontotemporal dementia, 13 individuals with semantic dementia, and 8 individuals with progressive nonfluent aphasia. They retrieved 41 features using an Automatic Speech Recognition (ASR) system, including speech rate, the mean and standard deviation of pause, vowel, and consonant length. Using a multilayered perceptron network, they were able to achieve an 88% classification accuracy for AD vs. healthy participants based on lexical and auditory data.

Luz et al. [81], in their recent research, extracted features based on a graph that encodes turn pattern and speech rate from the Carolina Conversations Collection [94], which includes recordings of interviews with dementia patients and healthy people. By using these features, they composed an additive logistic regression model that could distinguish speech between healthy subjects and dementia patients.

Karlekar et al. [95] presented a neural classification framework for detecting AD using the DementiaBank dataset. Their study applied CNNs, LSTMs, and a hybrid CNN-LSTM model to classify utterances from patients as either AD-positive or control. Notably, their best-performing model, the CNN-LSTM architecture trained on POS-tagged data, achieved an accuracy of 91.1%, setting a new benchmark at the time of publication.

However, a critical assessment of their methodology reveals several factors that likely contributed to the reported high accuracy:

1. **Dataset Imbalance.** The dataset used in their experiments is highly imbalanced: 1,017 transcripts from AD patients versus only 243 from control participants. After preprocessing, this yielded 11,458 AD-labeled utterances and 2,904 control utterances, resulting in a class distribution of approximately 80:20. A naive classifier that always predicts the majority class (AD) would already achieve a baseline accuracy of approximately 79.8%. This inflated chance level must be taken into account when interpreting their results.
2. **Train/Dev/Test Split Scheme.** The authors adopted a fixed 80/10/10 split for training, development, and testing, rather than employing cross-validation. While this is common in deep learning workflows, it may introduce variance due to dataset-specific biases, especially in the presence of class imbalance. Cross-validation would have provided a more robust measure of generalizability.
3. **Task-Specific Predictive Features.** Many of the utterances stem from structured tasks such as the “Cookie Theft” description and the “Recall Test.” These tasks elicit contextually narrow responses. For example, omission of expected nouns (e.g., “cookie”) or excessive use of filler words (e.g., “uh,” “um”) are often strong indicators of AD. As a result, classification under these conditions is likely to be easier due to the constrained and distinguishable nature of task-induced language.
4. **Elevated Chance-Level Accuracy.** Due to the data imbalance and constrained linguistic context, the effective chance level is already high (approximately 80%).

Thus, the 91.1% reported accuracy represents only a modest improvement over chance.

In summary, while the work by Karlekar et al. [95] represents an important step toward interpretable neural approaches for AD detection, the high accuracy reported must be viewed in the context of the data imbalance, task structure, and evaluation strategy. In our study, we re-implement their approach as a baseline on our dataset with the same data split scheme as used in this work. Other than the data split scheme, everything else strictly follows their work, including the model architecture, hyperparameters, and training procedures.

In some studies, signal processing, and NLP techniques are used to detect signs of dementia that may be imperceptible to human professionals. For example, Tóth et al. [96] discovered that even though human annotators could not recognize pauses (sounds like "hmmm," etc.) reliably, these features are easy to collect with an ASR system. In this research, several acoustic parameters (hesitation ratio, speech speed, length, number of silent and filled pauses, and duration of utterance) were extracted from the recorded speech of 38 healthy controls subjects and 48 patients with MCI talking about two short films. They found that ASR-extracted features outperformed manually computed features (69.1% accuracy) when combined with machine learning approaches, notably with a Random Forest classifier (75% accuracy). König et al. [97] employed a similar machine learning approaches and showed an accuracy of 79% when discriminating MCI individuals from healthy counterparts, 94% for AD vs. healthy, and 80% for MCI vs. AD. Their tests, on the other hand, were conducted on non-spontaneous speech data collected under controlled settings as part of a neuropsychological evaluation that also included mechanically transcribed text.

The idea of combining two perplexity values, one from a language model trained on speech samples of dementia, and one from a language model trained on speech samples of healthy, was proposed by Wankerl et al. [98]. Perplexity is used to estimate the fit between a probabilistic language model, and a sample of previously unseen text in the training.

The n -gram language model is a method widely used in processing speech or written language [99]. N -gram language models create probability density from training text data by calculating the frequencies of the word sequences. In the simplest uni-gram/1-gram language model, the sequence only contains one word. The model counts the words in the training data and assigns a probability to them. For a sentence S :

$$S = (w_1, w_2, \dots, w_k),$$

$w_1, w_2, \dots, w_k$ represent the 1st word, 2nd word, ..., k -th word in the sentence S . For any sentence in the test data, the model estimates its possibility of existence based on the training data. In the case of the uni-gram language model, the sentence probability $p(S)$ equals the product of each word's probability $\prod_{i=1}^k p(w_i)$.

The uni-gram language model cannot comprise any contextual information because it only gives the probability distribution of individual words. On the other hand, calculating the probability distribution of individual sentences leads to a unique probability. It might be hard for the model to make proper predictions for new unseen data. Therefore, the length of sequences is limited to a certain small number. A model that calculates the probability distribution of sequences composed of n words is called the n -gram language model. For example, the tri-gram language model calculates the probability distribution of sequences composed of 3 words. For a sentence $S = (w_1, w_2, \dots, w_k)$ of k words, when the sequence length is n , the probability is evaluated by [98]:

$$p(S) = \prod_{i=1}^k p(w_i | w_{i-n+1}, \dots, w_{i-1}). \quad (2.1)$$

In training the n -gram language model, each sentence is padded with special tokens to indicate the beginning and end of the sentence. The token also helps to calculate the probability of the first word in the sentence. In the case where the data is limited, many sequences in the test data may not appear in the training data. Therefore, it is important to introduce a smoothing method into the n -gram language models. Additive

smoothing [100] simply assigns a constant probability for the sequence that did not appear in the training data.

As the first step in the work [98], two language models are created from all the data from dementia subjects (LM_{dem}) and all the data from the healthy subjects (LM_{con}). These two models represent the typical speech patterns of dementia speech and non-dementia speech. To evaluate the speech of each participant in the data set, an additional model needs to be created so that the test data itself does not appear in the training data set. For subjects with more than one recording sample, their speech samples are entirely excluded from the data set. This is necessary because each individual might use similar verbalization or re-occurring phrases that are not common or universal to the entire data set which might distort the perplexity distribution.

In the work [98], a cross-validation approach is applied. For each subject, s in the dementia category, a tri-gram model LM_{-s} is created that takes all the speech samples as the training data except those that belong to the subject s . The perplexity p_{own} of every speech from s is calculated using LM_{-s} . While p_{other} is obtained using the model LM_{con} , which certainly does not contain any recording from subject s because they are in different categories. The cross-validation is repeated for all the subjects in the data set. In addition to p_{other} , p_{own} , the difference between them is added as another feature, which is calculated in the following:

$$p_{diff} = \begin{cases} p_{own} - p_{other}, & \text{if } s \in \text{AD Group} . \\ p_{other} - p_{own}, & \text{if } s \in \text{Control Group} . \end{cases} \quad (2.2)$$

By setting a threshold $p_{diff} = -1.41$, under equal error rate of both categories, they calculated the classification accuracy which equals 77.1%.

Fritsch et al. [101] improved the two perplexities methods by introducing neural networks to replace the n -gram language models in the original methods. Instead of conventional statistical language models, they trained LSTM neural network-based language models. They used the two perplexities methods with LSTM and achieved an accuracy of 85.6%.

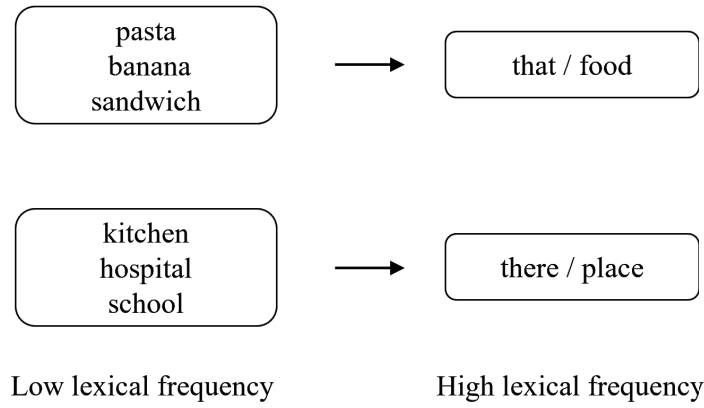


Figure 2.1. Examples of words with different lexical frequencies.

In 2020, Cohen et al. [102] interrogated the two perplexities methods by using artificially synthesized speech data that are created to simulate progressive dementia detection. Bird et al. [103] created synthetic narratives by creating a baseline sample and removing and/or replacing the nouns and verbs with higher lexical frequency (mother vs. woman vs. person). Lexical frequency shows how specific a word is in describing the context information. In Fig. 2.1, we give two groups of examples. In both examples, the words in the outer circles have broader meanings and that includes the meanings of the words shown in the inner circles. Cohen et al. [102] followed the work of Bird et al. [103] and implemented the two perplexities methods by comparing the original data (words are not replaced nor removed) and the modified data (some words are replaced with higher lexical frequency words). By doing so, they noticed that the perplexity distribution is highly influenced by words' lexical frequency. Their research confirmed that the lexical frequency of vocabulary is effective in detecting dementia.

Previous works analyzed language models and data. However, the analysis does not answer all questions about the topic. They found that language models' perplexities are associated with lexical frequency, but is it the primary information in the detection? Which one contributes the most to the neural network classifiers, syntax, or semantic aspects of the language? What kind of information composition or format do the neural networks take to improve the accuracy performance in dementia detection? Which part of the patients' language ability is impaired so the model thinks it is related to dementia? These questions are important for the development of more reliable, explainable, and

trustworthy models for dementia detection, as a black-box model, although it can perform dementia detection as well, does not meet the needs of the patients and clinicians, who may be more interested in understanding the symptom-level impairments rather than the model’s internal workings. Besides, when it comes to the medicine area, data is often limited and related to the personal privacy of the patients. Therefore, not all the data is available in the desired amount and form. This results in that deep learning often cannot reach its best performance. Hence, answering these questions helps us develop more reliable, explainable, and accurate models by manually manipulating the data we have.

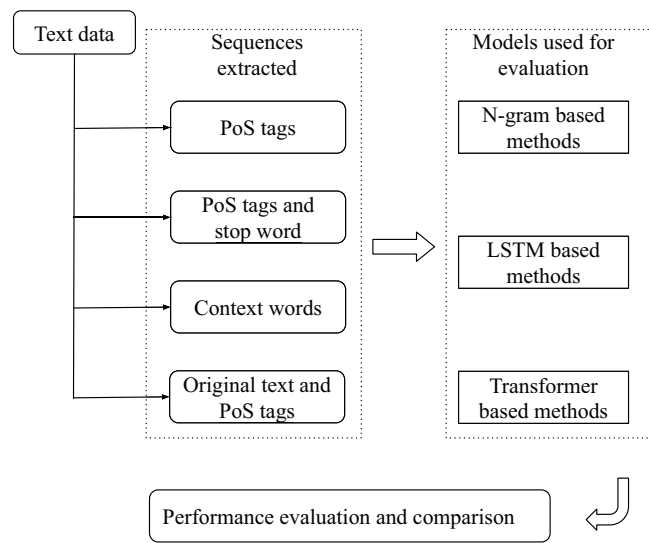


Figure 2.2. Overall pipeline of the research.

In this work, we first explored whether the richness, specificity, or variety of the vocabulary along with the difficulty to predict the next word should be the primary indicators in the task of dementia detection. Or if the text’s grammatical structure may be a better indicator of dementia detection. We re-implemented the two perplexities methods with PoS tags and stop word sequences. A PoS tag is a category to which a word is assigned according to its syntactic functions. For example, some simplified and regular PoS tags in English are verb, noun, adjective, etc. Stop words are words that do not contain contextual information, nor do they indicate the context or meaning of the sentences.

Using PoS tags reduces the vocabulary size, which helps to improve the generality of the n-gram language models. In the original methods, they used every word in the data set, but in our implementation, only 33 PoS tags and 127 stop words are used. This can

largely decrease the possibility of unknown sequences in the test data and increases the generality of the models. It enables the n-gram models to utilize the syntax of texts in an explicit way, especially in the case where the data set is limited. We also explored using fewer complex models to perform the classification task. Less complex models are easier to train and less likely to have overfitting problems.

Despite the use of PoS tags and stop words are common techniques in NLP, the novelty of our work lies in the combination of these two techniques to separate the semantic and syntactic components of the language. This separation allows us to evaluate the contribution of each component to dementia detection. By analyzing the performance of models trained on different information compositions, we can determine whether the semantic part (context words) or the syntactic part (PoS tags and stop words) is more significant in the task of dementia detection. This approach provides a new perspective on the role of language structure in cognitive health assessment and helps us understand the importance of different linguistic features in detecting dementia. In medical settings, post hoc explanation is crucial for the acceptance of AI-based diagnostic tools. By using PoS tags and stop words, we can provide insights into the decision-making process of our models.

Different from the interpretability in machine learning, which often focuses on model-specific features like attention weights, our approach emphasizes post hoc explanations of the symptoms and language impairments of the patients. Interpretability in machine learning refers to the ability to understand and explain a model’s decision-making process, typically in terms of feature contributions, statistical relevance, or algorithmic behavior. This differs markedly from how non-experts, such as a patient diagnosed with dementia, understands explanations. While machine learning interpretability focuses on quantitative outputs and feature-level analysis, such as the statistical weight of a memory score, patients generally expect intuitive, high-level, and clinically grounded justifications, framed in terms of symptoms, diagnoses, or treatment implications. Furthermore, ML interpretability serves technical objectives like model validation or regulatory compliance, whereas lay understanding seeks personal relevance, emotional reassurance, and actionable insight. This divergence in abstraction, language, and purpose underscores

the need for translation layers that can connect computational explanations with human-centered narratives.

This distinction aligns with findings by Tonekaboni et al. [61], who emphasize that explainability in clinical contexts must be attuned to the cognitive and practical needs of end users—namely, clinicians and, by extension, patients. Their study demonstrates that technical interpretability methods, such as feature attribution or instance-based explanations, often fail to meet the communicative and contextual expectations of clinical stakeholders. Instead, clinicians value explanations that align with clinical reasoning, patient-specific narratives, and actionable guidance. This highlights a critical translational gap between the internal logic of machine learning models and the interpretive frameworks used by non-experts. Therefore, post hoc explanations that prioritize symptoms, observable impairments, and disease trajectories are essential for making computational decisions intelligible and trustworthy to patients and caregivers, particularly in emotionally and cognitively sensitive domains like dementia care.

Our method enables the identification of specific language impairments, such as deficits in vocabulary usage or grammatical fluency, which are critical markers in dementia detection. This level of granularity is particularly valuable in medical contexts, as it provides actionable insights into the patient’s cognitive and linguistic challenges. By offering a clear and understandable explanation for the model’s decision, answering questions like “Why does the model think I have dementia?”, we foster trust and transparency with both patients and clinicians. This transparency not only enhances the credibility of AI-based diagnostic tools but also empowers clinicians to make more informed decisions regarding diagnosis and treatment. Furthermore, understanding the specific linguistic deficits can guide targeted therapeutic interventions, ultimately improving patient outcomes and advancing the integration of AI in clinical practice. Therefore, despite the common use of PoS tags and stop words in NLP, our approach provides a novel perspective on e-healthcare, especially, dementia detection applications by combining these techniques to separate semantic and syntactic components to provide a more clinically relevant and justifiable explanation of the model’s decisions.

To achieve this, we created multiple classifier models based on different classifier architectures and evaluated the performance using different information compositions. We separate syntax and semantic components from the sentences by incorporating the PoS tags and stop words in the pre-processing of the data set. We researched what's the most efficient information input for the machine learning classifiers. We implemented two main neural network architectures: an LSTM-based neural network classifier built from scratch, and another pre-trained language model-based classifier to do the classification. For both architectures, we train 3 different models:

- a model trained with only context words (semantic information).
- a model trained with sequences composed of PoS tags of the words, which contain the sentence patterns information (syntactic information) but no context information.
- a model trained with sequences composed of both stop words and PoS tags, which contain the sentence patterns information but with finer details.

Lastly, we discussed whether vocabulary variety or richness should be the primary indicator for dementia detection. The overall pipeline of the research is shown in Fig. 2.2. In the work [102], they evaluated why the two perplexities methods work. They found that perplexities of neural network models are associated with lexical frequencies. We researched more about this topic by the aforementioned methods and showed by experiments that despite the importance of the context words, they are not necessarily the most valuable indicators for dementia detection.

The major contributions of our work are summarized as follows:

- We re-implemented the two perplexities methods with PoS tags and stop word sequences, which decreases vocabulary size, reduces unseen sequences in the test data, and improves generality by focusing on syntactic components of language.
- We created multiple classifier models based on different architectures and information compositions, achieving a classification accuracy of 81.54% under 10-fold

cross-validation using full transcripts combined with syntactic features. Our analysis highlights the importance of both syntactic and semantic components in dementia detection.

- We provided a novel perspective on dementia detection by systematically separating semantic and syntactic components of language. Our experiments demonstrate that sentence patterns and grammatical fluency are critical indicators for dementia detection, offering clinically relevant and justifiable explanations for model decisions.

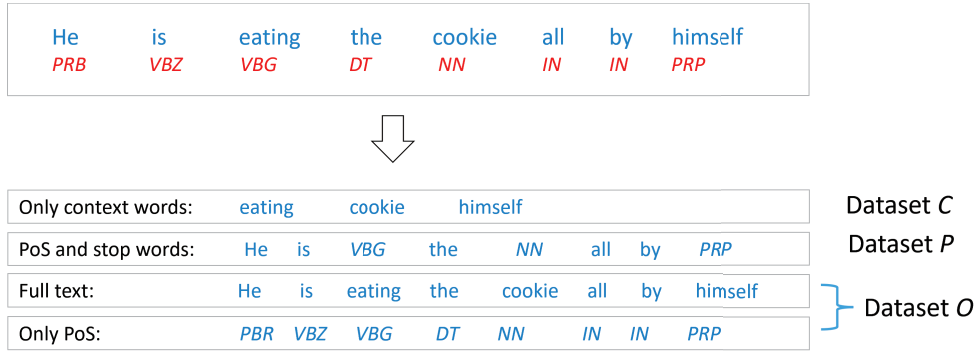


Figure 2.3. A flowchart of the pre-processing of the text data set. The original text is processed into 4 different instances that are used in the experiments.

2.3 Dementia Detection Based on Language Models

2.3.1 Data Set

The DementiaBank data set provided in TalkBank [104] is used in this work to evaluate the performance of the different introduced models in dementia detection. TalkBank is a multi-lingual data set established in 2002 whose object is to encourage research in human and animal communication. It is composed of data in several fields like first/second language acquisition, conversation analysis, dementia, etc. As one of its sub-sets, the DementiaBank is a set of video and audio data with their respective transcribed texts in different languages from dementia patients and non-dementia subjects. Specifically, we

use the Pitt Corpus, which includes recorded audio samples with their manually annotated transcripts in a picture description task. The patients were requested to see a picture and describe the content of the picture with the help of an interviewer. It is believed that spontaneous speech is rich in information that could indicate the mental status and cognitive functions of the subjects. With the Pitt Corpus data set, we could detect dementia by evaluating their speech data with machine learning techniques. In this work, we limited our use of the data set to the Pitt Corpus, which contains 309 recordings from 166 AD patients and 242 recordings from 94 healthy subjects. The average number of words is 91 in the AD patients' text samples and 97 in the healthy subjects' text samples.

2.3.2 Pre-processing of Data Set

Instead of using the original data set, we extracted 4 kinds of information from the original text. The text in the Pitt Corpus is processed with the natural language parsing tools as follows. A PoS tagger is utilized to extract the PoS tags of every word in the texts. By converting every word into its corresponding PoS tags, we extract the PoS tag sequences from the given text in the data set. Besides PoS tags, we also used stop words to help extract information. Stop words are words that do not add contextual information to the sentences, nor do they indicate the context or meaning of a sentence. Either because they are of little significance in expressing the conception (like prepositions, conjunctions, etc.), or they are words that frequently appear in the specific speech samples that they do not contribute to the classification of machine learning models. Following the fact mentioned above, for each speech sample, 4 instances that are composed of different information (*i.e.* original texts, PoS tag sequences, and stop words list), are created. These instances are generated as described below (Fig. 2.3).

- **An instance with only context words:** All of the words in this case are context words. As previously noted, this is a typical method for deleting "noisy" text parts and enhances classification in a variety of NLP applications, including sentiment analysis [4]. Previous studies, such as [39], have used this method in the field of

dementia and CI detection. The speech samples processed in this manner are used to create a data set we refer to as $\mathcal{C}$.

- **An instance without context words:** The context words in the speech samples are replaced with their PoS tags, yet we keep the stop words as they are. Although it is counterintuitive, we process the data in this way because it allows us to notice when a phrase or paragraph does not follow the natural flow of language and reveal common language patterns regardless of the context. Despite its lack of value in tasks such as sentiment analysis or hate speech detection, we believe that this information is highly useful when dealing with the issue of dementia diagnosis. The speech samples processed in this manner are used to create a data set we refer to as $\mathcal{P}$.

- **An instance where the words in the original text sample is coupled with their PoS tags:** In this case, we extracted the PoS tags of the original speech and create sequences of both. In training the network, both sequences (the text’s word sequence and the PoS tag sequence) are fed into our neural networks. It is worth mentioning that the PoS tag information gets included in the word representation in some way or another, whether using a pre-trained embedding matrix or when training one from scratch. However, because of the large size of the embedding matrix and the short length of the corpus, this information may be lost in both cases. The speech samples processed in this manner are used to create a data set we refer to as $\mathcal{O}$.

- **An instance with only PoS tags:** The instance is made up entirely of PoS tags with no contextual information in the speech. PoS tags are related to information about syntax or phrase patterns in human language. In the image description task, where the object and context are extremely well defined, dementia may be detected by assessing simply the patterns of PoS tags. The speech samples processed in this manner are used to create a data set we refer to it as $\mathcal{T}$.

2.3.3 Baseline Methods

In this chapter, we implemented the the work of Karlekar et al. [95] as our baseline method, which employs neural network models to detect Alzheimer’s disease using linguistic data. We followed their experiment settings as closely as possible. Specifically, we adopt their three model architectures: a CNN model, a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN), and a hybrid CNN-LSTM model. The structures and hyperparameters of these models strictly follow the original paper: All models were trained with the Adam optimizer, using a learning rate of $1e^{-4}$, with the Gradients norms clipped to 2.0.

For the CNN model, a two-dimensional convolutional architecture was applied with filter widths of 3, 4, and 5, each with 128 filters. Training was conducted with a batch size of 128, and dropout was applied at a rate of 0.80 to reduce overfitting. The LSTM model consisted of two stacked LSTM layers, each with 128 hidden units. A smaller batch size of 32 was used, and dropout was applied at a rate of 0.70. The combined CNN-LSTM architecture placed an LSTM layer on top of the CNN output. The CNN component employed 100 filters for each of the filter widths 3, 4, 5, and 6. Input embeddings were 300-dimensional. The LSTM layer had 300 hidden units. Both standard dropout on the CNN and recurrent dropout on the LSTM were set to 0.65.

Following Karlekar et al. [95], we conduct our experiments on both raw and POS-tagged text of the dataset. Evaluation is conducted using accuracy, precision, recall, and F1-score as metrics. Because the original paper conducted experiments on an extremely imbalanced dataset, and they did not provide the exact data split, we implemented this method based on the datasets and data splits we have used in our work. Specifically, we used the Pitt Corpus Cookie Theft dataset and split it into training and test sets using a Leave-One-Subject-Out (LOSO) cross-validation scheme as in the two-perplexity methods, and a 10-fold cross-validation as employed in the neural network-based methods.

2.3.4 N-gram Language Model and PoS Tag Based Dementia Detection

In their previous works [98, 102] on the two perplexities methods, they used the original transcripts with full vocabulary in their experimental setup. As previously stated, sequences made from PoS tags combined with stop words can represent the grammar and sentence patterns information while erasing all the contextual information that could indicate the actual concept in the language.

In the example shown in Fig. 2.3, the sentence "He is eating the cookie all by himself!" is converted into the sequence "He is VBG the NN all by PRP!" The sequence kept almost all the grammar information from the original sentence, even though the specific contextual information is hidden by the PoS tags. This sequence could help us evaluate how grammar and sentence patterns information is contributing to the two perplexities methods to detect dementia.

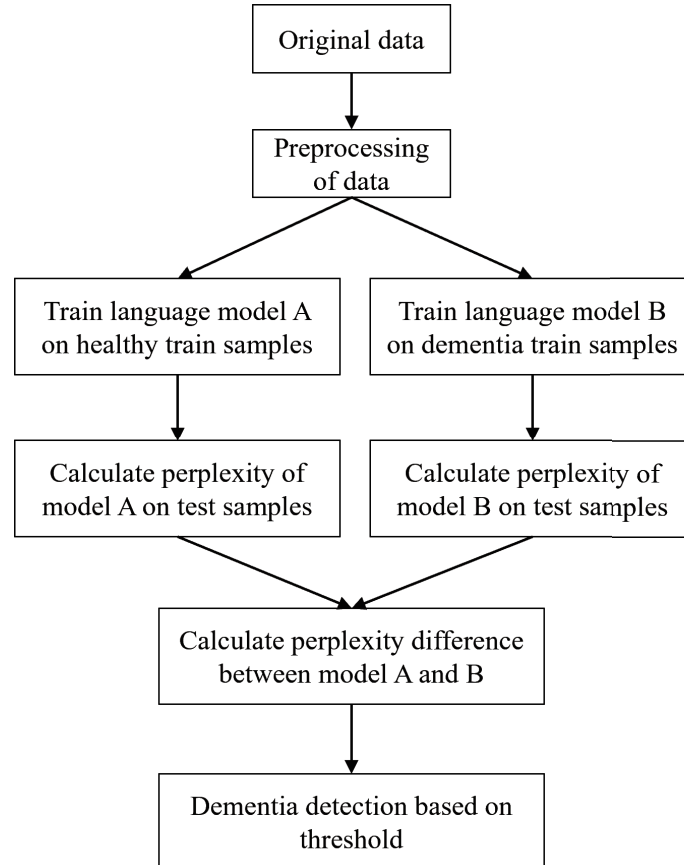


Figure 2.4. A flow chart of using two perplexities methods to detect dementia.

Under this idea, we implemented the two perplexities methods following the work of Wankerl, et al. [98]. However, we do not use the transcripts with full vocabulary to evaluate the performance. Instead, the data set $\mathcal{P}$ is used in the implementation. Data set $\mathcal{P}$ in which the sequence is only composed of PoS tags and stop words. We used the tri-gram language model, which would count the frequency distribution of 3-word sequences, in the implementation to perform the two perplexities methods. The tri-gram language model is a pure statistical language model. Thus, it does not have any capability to assign a value for the sequence that does not appear in the training data set. Yet, as the Pitt Corpus is a very limited data set containing only around 500 pieces of samples with less than 100 words in each, the occurrence of unseen sequences in the test data is likely to happen. Hence, a smoothing method that avoids this issue is necessary to ensure the system works well on the test data. Additive smoothing simply assigns a fixed value to the sequences that do not appear in the training data [100]. In this experiment, we use Laplace smoothing which is a kind of additive smoothing. In implementing the two perplexities methods, we evaluate the classification accuracy by using a Leave-One-Subject-Out (LOSO) scheme. Each time, we hold all the samples of one subject out as test data and use the rest as training data. This is because, in the Pitt Corpus data set, many subjects visit the doctor more than one time, and they contribute more than one speech sample to the data set. While describing the cookie theft figure, speech samples of the same subject may use similar sentence patterns or vocabulary to describe the figure. Thus, to prevent the model from learning the identity information of the subject, we hold all samples of one subject out each time. Following the two perplexities methods of Wankerl, et al. [98], we train two language models on dementia data and non-dementia data respectively and use the two language models to calculate the perplexities of the test samples. By checking the perplexity difference between the two language models, we decide if the test data belongs to which category by setting a threshold. The flowchart of this method is shown in Fig. 2.4.

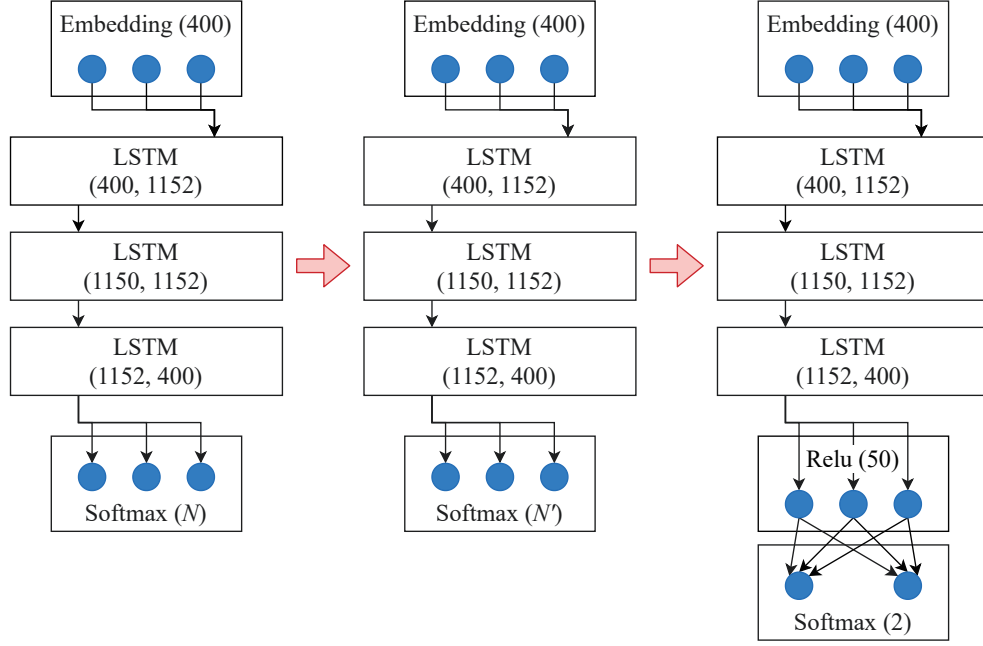


Figure 2.5. The AWD-LSTM neural network architecture.

2.3.5 Neural Network-Based Dementia Detection

We employ two neural networks for classification as previously described. The first neural network is based on the Averaged Stochastic Gradient Descent Weight-Dropped LSTM (AWD-LSTM) as proposed in [105]. The second neural network is trained from scratch using a standard attention network architecture. We will show the details of these networks in the following part.

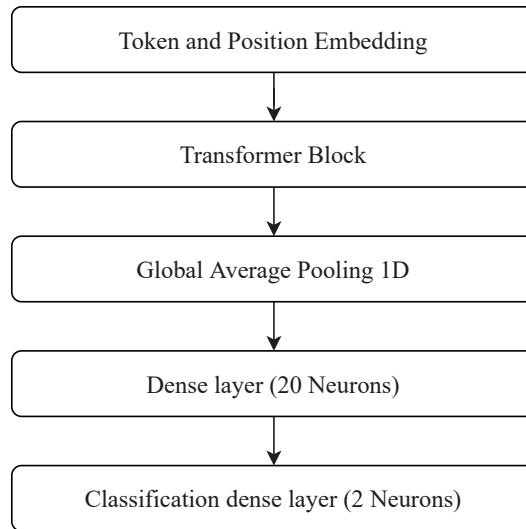


Figure 2.6. The transformer neural network architecture.

Pre-trained AWD-LSTM Network

Original Language Model. We make use of Howard and Ruder’s pre-trained language model, which is explained thoroughly in [106]. This language model was trained using WikiText-103 [107], a data set consisting of 28 595 pre-processed Wikipedia articles with about 103 million different words. Howard and Ruder proposed AWD-LSTM in their work. AWD-LSTM is a standard LSTM network with variably adjustable dropout hyperparameters at its core. The AWD construction is shown in the leftmost half of Fig. 2.5. As shown in the figure, after the embedding layer, there are three stacked LSTM layers followed by the prediction layer with a Softmax activation. The network’s overall number of parameters is manageable with an embedding size of 400 and 1152 activations per LSTM layer.

Target Task Language Model Fine-Tuning. We fine-tune the pre-trained language model using the dementia data set in hand. We use all of the data in the data set. In this stage, the labels are not utilized at all because our aim at this point is for the language model to learn to understand specific linguistic characteristics. Linguistic characteristics here refer to how words are related to one another and the hidden meanings of slangs, etc. The size of the embedding matrix model is $N' \times 400$, where N' represents the number of different words in the data set, which is also the size of the first layer of the network. To fine-tune the model, the Universal Language Model Fine-Tuning (ULMFiT) technique proposed in [105] is employed, which involves progressively unfreezing and adjusting learning rates. The softmax layer is the first to be unfrozen, enabling its parameters to be fine-tuned using the learning rate of 0.1 for the first 1 epoch. The remaining layers are then unfrozen and adjusted with a learning rate of 0.001 for 5 epochs, after which we continue training by lowering the learning rate. The dropout rate for all the layers is set to 0.3, while the Adam optimizer’s parameters β_1 and β_2 are set to 0.90 and 0.99, respectively.

Target Task Classifier. The last step in the language model adjustment process is the classification. We used two linear blocks, one with Rectified Linear Unit (ReLU) activation and the other one with softmax activation to substitute the last softmax layer in

the original network. In other words, in this model, they are inserted after the three LSTM layers. This is because the model is no longer employed as a language model to predict the next word, but as a classification model to predict the text’s class. The model is fine-tuned using slow unfreezing, discriminative learning rates, and slanted triangle learning rates. Unlike the previous phase, we utilize the training set to fine-tune the model (while leaving the test examples to classify out) and execute the classification on the test samples left out. The Softmax and ReLU layers are unfrozen first, then fine-tuned for 1 iteration with a learning rate of 0.01. After that, we unfreeze the third LSTM besides the Softmax and ReLU layer and fine-tune the whole system to its optimal values. For both the dense layers with Softmax and ReLU activations, the learning rates are set at 0.05. The learning rate of the LSTM layer is set by following the work [106], which states that if the final layer’s learning rate is η^l , prior layers should have a learning rate of $\eta^l - 1 = \eta^l/2.6$. Similarly, we unfreeze the second LSTM layer next following the same rule but with a smaller learning rate. Finally, the whole network is unfrozen and trained using the previously mentioned progressively decreasing learning rate described above.

Attention-Based Network

We implemented a Transformer-based network architecture in the second model for classification. By replacing LSTM layers with what is known as an attention mechanism [108], the transformer network solved many issues of LSTM training difficulties. The network is made up of an embedding layer that encodes the token, its position, and segmentation information, and an attention-based transformer block, followed by 2 dense layers where the second dense layer is used for classification. The architecture of the transformer-based network is given in Fig. 2.6.

The main differences between the proposed methods and the previous works are summarized in the Table 2.1.

Table 2.1. Comparison between our research and the referred research.

Category	Ours	Wankerl [98]	Cohen [102]
Pre-processing	Pre-processing with PoS tagger and stop words	Used the original transcripts	Used the original transcripts
Classifiers	N-gram, LSTM, and Transformer language models	N-gram language models	LSTM language models
Information source	Syntax information or semantic information or both	Full transcript	Full transcript
Research methods	Analysis with extracted individual information components	Classification	Analysis with artificially synthesized data

2.4 Evaluation of Performance

2.4.1 Baseline Methods

For the baseline method, we strictly followed the experiment settings as described in [95] and implemented the three neural network architectures on the same data set and data split as in our work to make a fair comparison.

Model	Accuracy	Precision	Recall	F1 Score
CNN	0.7717	0.7868	0.8123	0.7994
LSTM	0.6938	0.7188	0.7443	0.7313
CNN-LSTM	0.7029	0.7164	0.7767	0.7453
CNN-LSTM / PoS tagged	0.6250	0.7361	0.5146	0.6057

Table 2.2. The performance of the experiments using CNN, LSTM, and CNN-LSTM models on the Pitt Corpus data set under 10-fold cross-validation. The results are reported as accuracy, precision, recall, and F1 score.

Tables 2.2 and 2.3 show the performance of the three neural network architectures on the Pitt Corpus data set under 10-fold cross-validation and Leave-One-Subject-Out (LOSO) cross-validation, respectively. The results are reported as accuracy, precision, recall, and F1 score. The CNN model achieved the highest accuracy of 0.7717 under 10-fold

Model	Accuracy	Precision	Recall	F1 Score
CNN	0.7446	0.7515	0.8123	0.7807
LSTM	0.7138	0.7428	0.7476	0.7452
CNN-LSTM	0.7283	0.7345	0.8058	0.7685
CNN-LSTM / PoS tagged	0.6685	0.7405	0.6278	0.6795

Table 2.3. The performance of the experiments using CNN, LSTM, and CNN-LSTM models on the Pitt Corpus data set under Leave-One-Subject-Out (LOSO) cross-validation. The results are reported as accuracy, precision, recall, and F1 score.

cross-validation and 0.7446 under LOSO cross-validation. The LSTM model achieved an accuracy of 0.6938 under 10-fold cross-validation and 0.7138 under LOSO cross-validation. The CNN-LSTM model achieved an accuracy of 0.7029 under 10-fold cross-validation and 0.7283 under LOSO cross-validation. The CNN-LSTM model with PoS tags achieved the lowest accuracy of 0.6250 under 10-fold cross-validation and 0.6685 under LOSO cross-validation. The CNN model outperforms the LSTM and CNN-LSTM models in terms of accuracy. This results may indicate that the CNN model is more effective in capturing the linguistic features of the data set when models are not finetuned based on language models but trained from scratch.

2.4.2 PoS Sequences based Classification

We first implemented the two perplexities methods using two tri-gram language models on the data set instance $\mathcal{T}$, where the sequences were made up entirely of pure PoS, tags to perform the two perplexities techniques. The perplexity difference results are presented in Fig. 2.7. The y-axis of each point in the figure shows the subject’s perplexity difference between the two language models. The x-axis is the patient index manually assigned to each subject. For the held-out test subject, each test subject contributes multiple samples because they visit the doctor more than one time during their treatment. There are two optional schemes for the machine learning model to perform classification. It can decide the category for each speech sample individually. It can also decide the category of the patient after concatenating the samples of him or her together as a big sample.

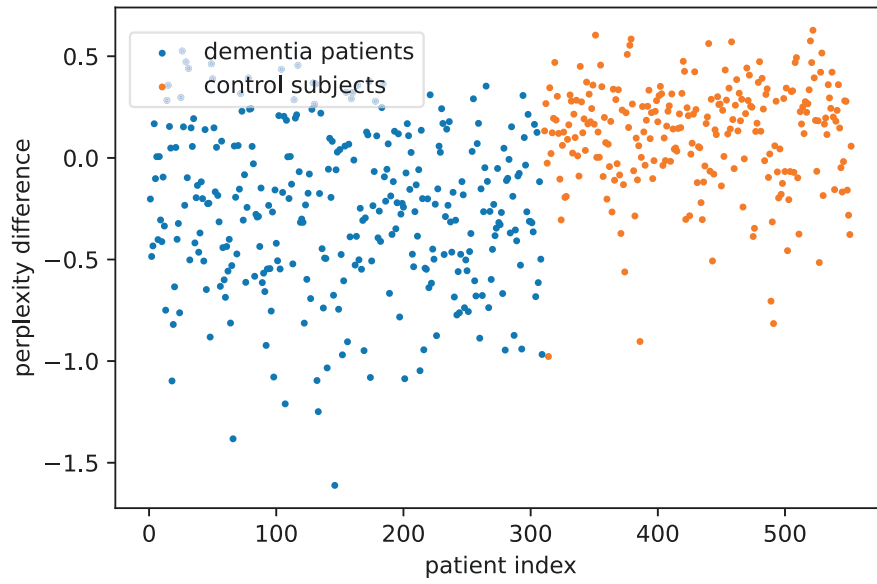


Figure 2.7. The perplexities difference between language models using only Part-of-Speech tags.

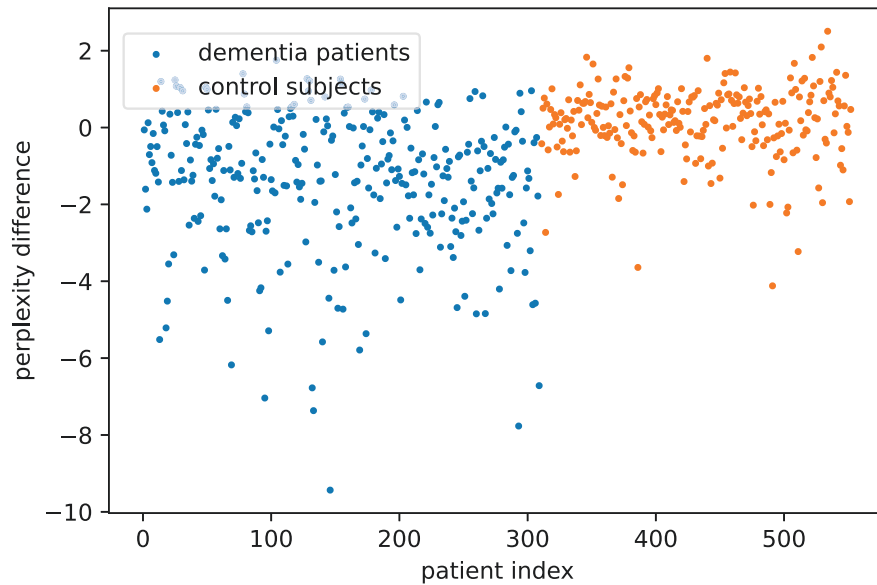


Figure 2.8. The perplexities difference between language models using Part-of-Speech tags and stop words.

We determine if a test sample is a dementia patient or a healthy control subject by setting a threshold for the perplexity difference. We reached an accuracy of 75.3% in the task of classification for each patient. We achieved a classification accuracy of 71.5% in the task of classification for each sample using an equal error rate as the threshold.

We implemented the two perplexities methods using two tri-gram language models on the data set instance $\mathcal{P}$, where the context words are replaced by PoS tags and the stop words are kept as they are. The perplexity difference results are presented in Fig. 2.8. The range of the Y-axis is different from the previous one because we add stop words in the training process and the vocabulary for training is different, which makes the complexity of the models different.

We reached an accuracy of 80.8% in the task of classification for each patient, and 72.8% in the task of classification for each sample using an equal error rate as the threshold. The accuracy performance improved when the stop words are included, approaching that of [98]. The Receiver Operating Characteristic (ROC) curve of this experiment is also shown in Fig. 2.9. The ROC curve shows the true positive rate and false positive rate of our method under different thresholds. It shows the trade-off between sensitivity and specificity. We achieved an AUC of 0.78. To diagnose dementia, we utilized only 36 PoS tags with 127 stop words. It drastically lowers the cost of calculation and annotation while maintaining high accuracy.

In this research, we evaluated the accuracy based on the strategy of Equal Error Rate (EER) to determine the threshold, following the work of Wankerl et al. [98]. This is because the EER is a common method to determine the threshold in the two perplexities methods. Besides, using EER as the threshold also makes it easier to compare the results with the previous works. However, it is worth mentioning that the threshold can be set differently based on the application scenario. People can prefer different thresholds based on the application scenario, such as the cost of false positive and false negative. In the medical field, we have investigated several researches that explored the preference of the patients and the doctors. For example, Deneef et al. [109] investigated the preferred diagnostic strategy of the doctors. These doctors can identify his or her own preferred

diagnostic strategy by answering the question: “How many additional treatment errors am I willing to make in order to treat one additional person correctly?”

Deneef et al. [109] introduced the use of a *marginal tradeoff ratio*, denoted as R , to quantify diagnostic decision preferences. This ratio represents the number of additional treatment errors (false positives) a decision-maker is willing to accept in order to achieve one additional correct treatment (true positive). Formally, the ratio is defined as:

$$R = \frac{H}{B}$$

where H is the net harm of treating a non-diseased patient (false positive), and B is the net benefit of correctly treating a diseased patient (true positive). Within the framework of expected-utility theory, R corresponds to the slope of the Receiver Operating Characteristic (ROC) curve at the optimal operating point. This slope can be used to select an appropriate threshold that aligns with a decision-maker’s tolerance for error tradeoffs.

To assess physician preferences, Deneef et al. conducted a survey in which respondents were asked:

“How many additional treatment errors are you willing to make in order to treat one additional person correctly?”

Their results indicated that most physicians preferred R -values between 10 and 20. This suggests that physicians are generally willing to accept a 1 to 2 wrong treatments (false positives) for every correct treatment (true positive) they administer. This preference reflects a cautious approach, prioritizing the minimization of overtreatment

In contrast, previous analyses based on *patient utilities* suggest that patients often prefer more aggressive treatment approaches. These analyses imply much higher acceptable tradeoff ratios (e.g., $R \approx 800$, which means they accept 8 wrong treatments to find 1 correct treatment), favoring strategies that minimize the chance of missing a true case, even at the cost of more false positives. For example, patients’ preferences are aligned with treating nearly all suspected cases, regardless of test outcomes. A comparison of preferences is summarized in Table 2.4.

Table 2.4. Comparison of Physician and Patient Preferences in Diagnostic Tradeoff

Aspect	Physicians	Patients
Preferred tradeoff ratio R	10–20	≈ 800
Preferred strategy	selective testing	Treat most or all suspected cases
Emphasis	Avoid overtreatment	Avoid undertreatment

These findings underscore the importance of tailoring diagnostic thresholds to specific application contexts and stakeholder preferences. In systems where thresholds are set automatically (e.g., using Equal Error Rate), it is important to consider whether the implicit tradeoff aligns with the operational objectives and user values.

Another research conducted by Sommerlad et al. [110] reported the sensitivity and specificity of Alzheimer’s disease (AD) diagnosis strategy currently practiced in the UK. Their paper reported that the sensitivity of the AD diagnosis strategy is around 50% to 60%, while the specificity is around 90% to 95%. This also indicates that the, in practice, the physicians prefer to avoid overtreatment, which is consistent with the findings of Deneef et al. [109].

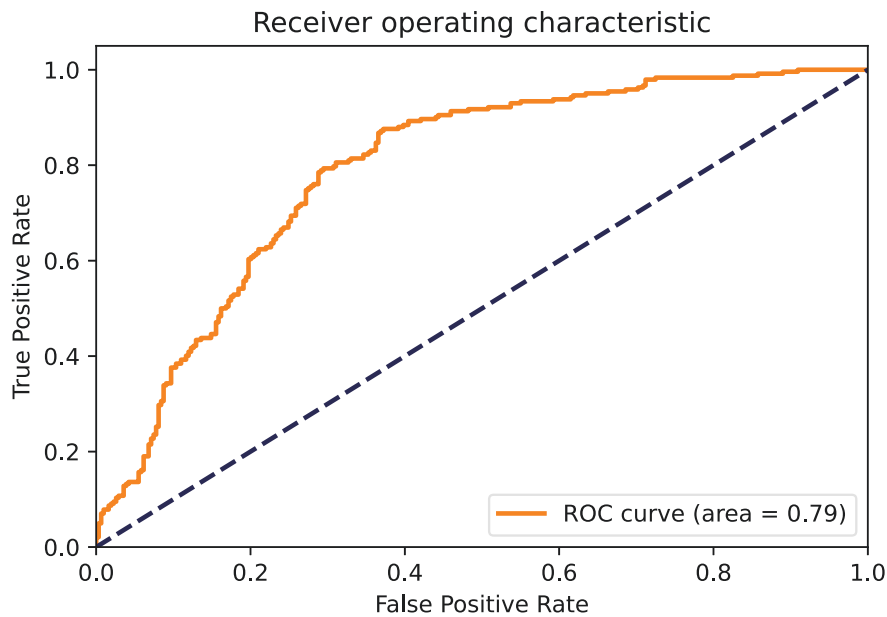


Figure 2.9. Receiver Operating Characteristic curve of classification performance for using sequences of PoS tags and stop words.

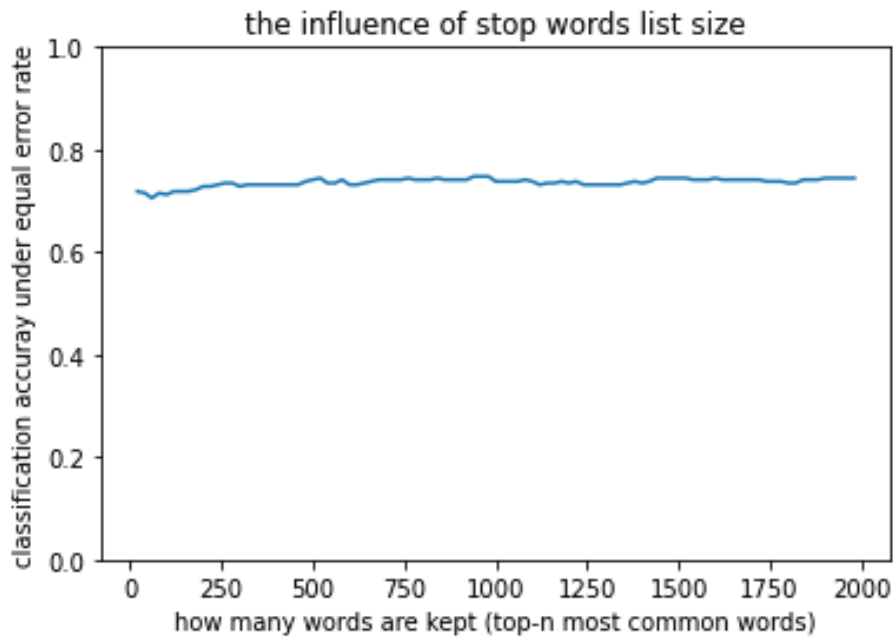


Figure 2.10. The influence of stop words list size: from 20 most used words to 2000 most used words.

In addition, we implemented experiments to see how much influence the high-frequency words have on the two perplexities methods. In the previous experiments, we used default stop words in the Natural Language Toolkit (NLTK) [111] as the high-frequency words to keep, because they do not have any contextual information. However, despite being on top of the list of the most commonly used words in English, stop words are not necessarily the only ones on this list. In the following experiments, we created the full list of the top English words according to common corpora, where each word is ranked based on its appearance frequency. Using this list, we pick the top N words to form the actual lists of words to keep. We composed the lists that we are going to keep strictly following the order of English word frequency. In the experiments, we create 100 sub-lists following this logic. The 1st list includes the 20 most used words in English. The 2nd list includes the 40 most used words in English. The 3rd list includes the 60 most used words in English. The 100th list includes the 2000 most used words in English. For each of the lists, we re-implement the two perplexities method and see how the accuracy behaves. The results are shown in Fig. 2.10. When the size of kept words list is less than 260, the results demonstrate that increasing the size of the kept words list improves

the classification accuracy moderately. Increasing the size of the list once it reaches beyond 260 does not improve the classification accuracy. We discovered that increasing the size of the kept words list does not always promise an increase in the performance of the n -gram language model based on the two perplexities technique. To identify dementia, we assume that the n -gram language model captures particular sentence patterns from the sequences of PoS tags. Dementia symptoms are reflected not just in vocabulary choice, but also in the syntax and sentence structures.

Furthermore, the word employed is heavily influenced by the data collection. Subjects in the Pitt Corpus data collection are requested to complete a picture description task. As a result, the number of words utilized is restricted to those contents that are visible in the image. When the subject is asked to explain what he or she is seeing, the vocabulary available to him or her has already been limited to a tiny number of terms related to the image. This implies that, regardless of how large the stop words list is, the majority of terms are not utilized in the description. Due to the small size of the Pitt Corpus data set, training an n -gram language model on PoS tags with stop words saves time and prevents over-fitting. It merely handles 36 PoS tags and a few stop words on the one hand. Converting words into their PoS tags, on the other hand, reduces the likelihood of seeing unexpected words/sequences in the test data, which helps minimize over-fitting.

2.4.3 AWD-LSTM Network

We display the performance of the classification using the AWD-LSTM network in Table 2.6. When utilizing the original text as is (data set $\mathcal{O}$), the classification accuracy, recall, and F1 score approach 81.54%, 83.13%, and 81.59%, respectively. When only context terms are used (data set $\mathcal{C}$), these percentages decrease to 66.54%, 66.87%, and 71.18%, respectively. This emphasizes the fact that some of the categorization information is incorporated in the full text and is lost when just context words are utilized. To demonstrate this point, we look at classification results using only stop words and replacing context words with their PoS tags. Precision, recall, and F1 score are all 76.15 %, 76.51%, and

80.38% in this example, respectively. Using the data set instance $\mathcal{P}$ shows a lower classification performance. Despite not utilizing any of the context words to train the network, the findings are rather good, leading us to assume that vocabulary richness and variety are not the only factors that may be utilized to detect dementia. The sentence’s grammatical structure is a solid information source for categorization. Compared to the baseline methods, the AWD-LSTM network achieves a higher accuracy of 81.54%, while the baseline methods achieved an accuracy of 77.17% under the same data split and data set.

2.4.4 Attention-Based Network

The performance of the classification using the Attention-based network on some of the instances discussed above is shown in Table 2.7. A similar phenomenon can be observed as can be seen in the experiment when we are using the AWD-LSTM networks. The classification accuracy, recall, and F1 score are 78.46%, 80.72%, and 82.72%, respectively, while utilizing the original texts (data set instance $\mathcal{O}$). When the stop words are removed and just context words are used (when utilizing the data set $\mathcal{C}$), these KPIs fall to 70.00%, 68.47%, and 74.51%, respectively. Finally, these KPIs reach 73.08%, 72.89%, and 77.56%, respectively, when the context words are substituted by their PoS tags (when utilizing the data set $\mathcal{P}$). This is consistent with our findings when using the AWD-LSTM network. Using the data set instance $\mathcal{O}$ gives us the highest accuracy. However, considering the situation when only context words are used (instance $\mathcal{C}$), or when only non-context words are used (instance $\mathcal{O}$), the latter gives us higher classification accuracy. This is because the sentence structure is a solid information source and it is mostly conveyed by sequences of PoS tags and stop words. Compared to the baseline methods, the Attention-based network achieves a relatively higher accuracy of 78.46%, while the baseline methods achieved an accuracy of 77.17% under the same data split and data set.

2.4.5 Discussion

Syntax, phonology, morphology, semantics, and pragmatics are the four major areas of linguistics. Syntax is concerned with how words are put together to form constituents

(words and sentences), and then how those constituents are placed in a certain sequence to communicate meaning. In other words, it focuses on the form of sentences and what constitutes a valid sentence. The study of how linguistic utterances and their meanings are connected, as well as how context impacts meaning, is known as semantics and pragmatics. While semantics and pragmatics have traditionally been employed extensively in NLP tasks involving text categorization, the syntactic elements that may be extracted from the transcribed texts of dementia patients may be more useful in the case of dementia diagnosis.

Table 2.5. Classification Performance of the n -gram language models on the 2 Sets $\mathcal{P}$ and $\mathcal{T}$.

Set	Accuracy	Precision	Recall	F1 score
Data set $\mathcal{T}$	65.15%	63.27%	90.29%	74.40%
Data set $\mathcal{P}$	72.78%	77.32%	72.82%	75.00%

In Table 2.5, we show that the n -gram language model-based two perplexities methods can detect dementia with an accuracy of 72.78% without any context words. Fig. 2.10 shows that keeping more high-frequency words could improve the performance of the system. Yet, after the number of kept words reaches beyond 260, keeping more words does not necessarily improve the performance. We believe that this is because, among the most used English words, words without context information contribute to the proper grammar structure account for the main part. By keeping these words in the data, more grammar information could be conveyed by the sequence of PoS tags and kept words. After the number of kept words goes beyond 260, the later included words are mostly conveying specific context information but they do not add much information in terms of grammar.

Table 2.6. Classification Performance of the AWD-LSTM Model on the 3 Sets $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{O}$.

Set	Accuracy	Precision	Recall	F1 score
Data set $\mathcal{C}$	66.54%	77.62%	66.87%	71.18%
Data set $\mathcal{P}$	76.15%	84.67%	76.51%	80.38%
Data set $\mathcal{O}$	81.54%	87.34%	83.13%	85.19%

Table 2.7. Classification Performance of the Attention-Based Model on the 3 Sets $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{O}$.

Set	Accuracy	Precision	Recall	F1 score
Data set $\mathcal{C}$	70.00%	81.43%	68.67%	74.51%
Data set $\mathcal{P}$	73.08%	82.88%	72.89%	77.56%
Data set $\mathcal{O}$	78.46%	84.81%	80.72%	82.72%

Table 2.8. Comparison of classification accuracy and interpretability approaches.

Methods	Accuracy	Comment
Wankerl et al. [98]	77%	
Fritsch et al. [101]	85%	
Cohen et al. [102]	87%	investigated impact of lexical frequency
Farzana et al. [89]	77%	domain-adaptive
Karlekar et al. [95]	77%	reimplemented on the same dataset and split
Ours	82%	post hoc explanations

Table 2.8 summarizes the classification accuracy of our proposed methods and compares it with previous works. Our method achieves an accuracy of 82% using the AWD-LSTM model with full transcripts, which is competitive with other methods. Despite not achieving the highest accuracy, our approach emphasizes the importance of syntactic components in language processing. By decomposing language into smaller components, we enhance model’s potential to provide justifiable and validatable post hoc explanations, which is crucial in medical settings. Our work is the first to systematically separate and evaluate the contributions of syntax and semantics for dementia detection using both statistical and neural models. This provides new insights into which linguistic features are most informative for diagnosis, and offers a practical path toward more transparent and explainable AI systems in healthcare. Our methodology enables clinicians to better understand model decisions, supporting trust and adoption in real-world applications.

2.4.6 Error Analysis

In the experiments of using n-gram language model based two perplexities methods, we achieved an accuracy of 80.8% in the task of classification for each patient, with a p-value

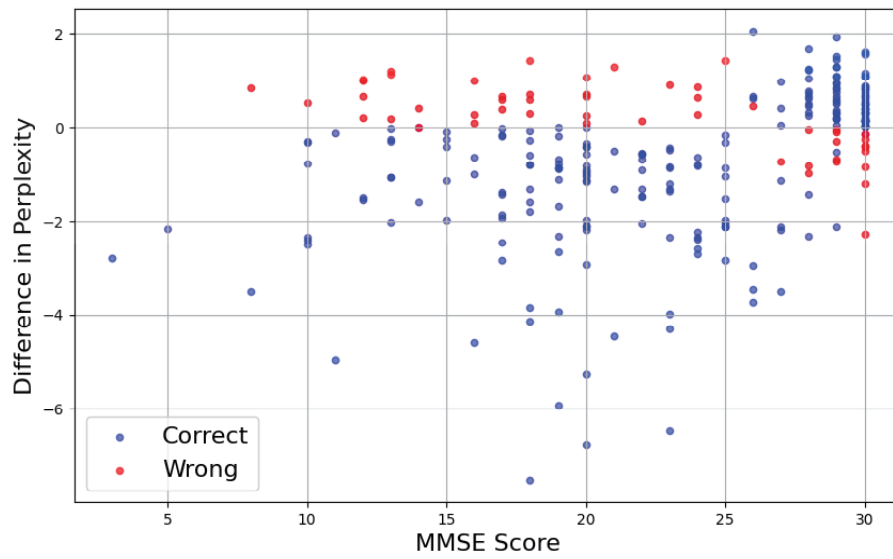


Figure 2.11. Difference in perplexity by MMSE score. Blue points represent correctly classified samples; red points are misclassified.

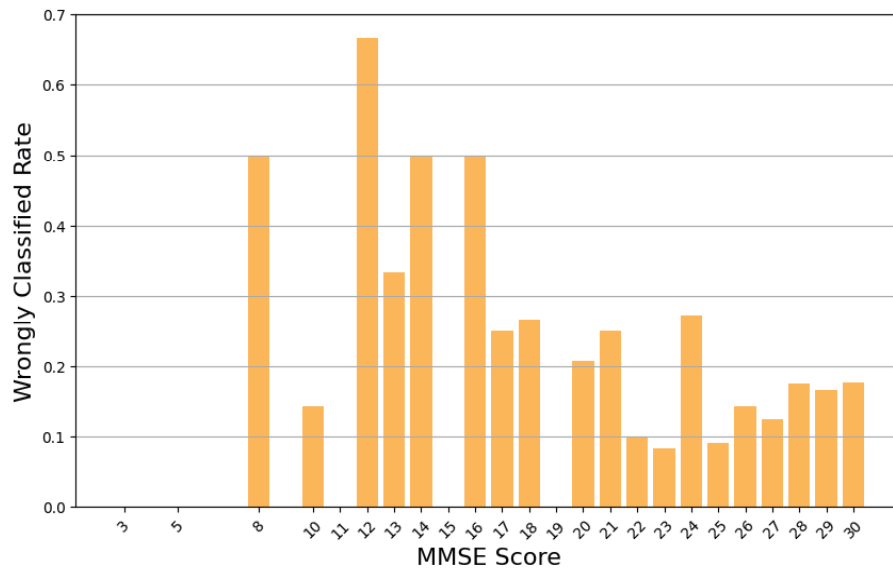


Figure 2.12. Misclassification rate by MMSE score. Patients with moderate MMSE scores have higher error rates.

of 8.02×10^{-22} . The precision, recall, and F1 score are 67.52%, 80.61%, and 73.49%, respectively. This indicates that the model is able to classify dementia patients and healthy controls with a high degree of accuracy. Comparing with wankerl et al. [98] and Farzana et al. [89], our method achieves a higher accuracy of 80.8% compared to their 77% with a p-value of 0.07, with the H_0 hypothesis that the two methods are equal under the binomial test. Although the threshold for significance is usually set at 0.05, we consider the p-value of 0.07 to be close enough to indicate a meaningful difference in performance considering the small size of the data set and the inherent variability in classification tasks.

To understand how the model’s performance varies with cognitive ability, we analyzed results by MMSE (Mini-Mental State Examination) scores. MMSE is a standard scale for assessing cognitive impairment, where scores closer to 30 indicate normal function and lower scores indicate more severe impairment. MMSE is a widely used cognitive assessment tool that evaluates various cognitive functions, including orientation, attention, memory, language, and visuospatial skills. The MMSE score ranges from 0 to 30, with higher scores indicating better cognitive function. Usually, a healthy subject has a score of 27 or above, while a score below 24 is typically indicative of cognitive impairment. Figure 2.12 shows the misclassification rate at each MMSE score. We observe that patients with scores between 11 and 17 have the highest error rates, in some cases above 60%. In contrast, participants with very low (severe dementia) or very high (no dementia) MMSE scores are more accurately classified. This suggests that patients with moderate impairment are harder to classify. Their language may include both normal and impaired patterns, making it more difficult for the model to assign a clear label. On the other hand, language from individuals with mild or severe impairment is more consistent with one class and easier to detect. Figure 2.11 shows the difference in perplexity between two language models for each patient, plotted against their MMSE score. Blue points represent correctly classified samples; red points are misclassified. These results indicate that classification is most difficult for patients with moderate cognitive decline. Their language is less distinct, and both models may give similar predictions. Future work could improve performance in this group by including more contextual or patient history information.

2.5 Conclusions

In this Chapter, we interrogated the role of language components in dementia detection by implementing two perplexities methods based on n -gram language models and two neural network architectures: AWD-LSTM and Attention-based networks. Unlike previous studies, in which stop words are removed from the text, we investigated the possibility of using the stop words themselves, since they provide non-contextual information that might aid in the detection of dementia. For this, we created three neural network models: one that solely processes context words, one that stops words with patterns of PoS tag sequences, and one that combines the two. We also implemented the two perplexities methods based on n -gram language models with different information sources. We demonstrate that both grammar and vocabulary contribute equally to categorization via experiments: the first model achieves an accuracy of 70.00%, the second model achieves an accuracy of 76.15%, and the third model achieves an accuracy of 81.54% under 10-fold cross-validation. The n -gram based two perplexities methods achieve an accuracy of 72.78% under LOSO cross-validation on sample classification and 80.00% on subject classification. Our results indicate that the information encoded in the text structure and the grammatical structure of the sentences have a larger role in categorization than the context itself.

Chapter 3

Dementia Detection Based on Anomaly Detection from Text Embeddings using Autoencoders

3.1 Introduction

Dementia detection from speech is a crucial task in the field of healthcare, as it can help identify individuals at risk of developing dementia and provide early intervention. Recent advancements in NLP and large language models have shown promising results in many NLP tasks [112], where the data amount is large enough to train the models. Large language models, such as BERT [92], GPT-3 [113], and LLaMA [114], have achieved state-of-the-art performance in various NLP tasks, including text classification, sentiment analysis, and question answering. These models are pre-trained on large corpora and can extract rich linguistic representations from text data, making them suitable for downstream tasks such as dementia detection.

Large language models have shown state-of-the-art performance in various NLP tasks. Compared to traditional machine learning models, they can capture more complex patterns and relationships in the data, which can be beneficial for tasks like dementia detection. Most importantly, large language models have the potential to make reasonable post hoc explanations on why the speech is abnormal, which is crucial for clinical applications for patients and doctors. Therefore, despite the fact that some works have achieved high accuracy in dementia detection, we believe that large language models can still provide valuable insights other than simply improving the performance of dementia detection from speech. However, the application of large language models to dementia detection is not straightforward. Most existing large language models are trained on large datasets, which may not be available in the context of dementia detection. The data for dementia detection is often limited, and the models trained on large datasets may not generalize well to this specific task. Therefore, in this chapter, we propose a new approach that leverages the capabilities of large language models to extract features from speech data and detect dementia using anomaly detection techniques to alleviate the data scarcity problems.

Pitt Corpus [115] is a commonly used dataset for dementia detection, but it contains only 552 speech samples with each sample being a short speech segment. This limited

amount of data poses a challenge for training large language models, which typically require large datasets to achieve good performance. Besides, the difference in the nature of dementia detection and other NLP tasks, such as sentiment analysis or question answering, makes it difficult to directly apply existing large language models to dementia detection. Other NLP tasks mainly focus on the semantic parts of the text, with the goal of understanding the meaning of the text. In contrast, dementia detection requires understanding the cognitive and linguistic features of the speech, such as fluency, coherence, and vocabulary usage. In the context of dementia detection, because the subjects are asked to describe the same picture, the focus is not on the semantic meaning of the speech (which is always equal) but rather on the structure and patterns of language use. Instead of understanding the meaning of the text, the goal is to identify deviations from normal speech patterns that may indicate cognitive impairment, which is aligned with the goal of anomaly detection.

Anomaly detection is a technique used to identify patterns in data that do not conform to expected behavior [116]. It involves identifying data points, events, or patterns that deviate significantly from the expected behavior within a dataset. These deviations, or anomalies, may indicate important underlying processes such as errors, fraudulent activity, structural defects, or novel phenomena, depending on the context of the application. In the context of dementia detection, anomaly detection can be used to identify speech patterns that deviate from those of healthy individuals. Anomaly detection methods can be particularly useful when the data is limited, as they do not require a large amount of labeled data to train the models. Anomaly detection models can be established using unsupervised learning techniques, where the model learns to identify normal patterns in the data and flag deviations as anomalies [117].

Motivated by the limitations of existing methods and the potential of large language models, we propose a new approach for dementia detection based on anomaly detection. In this chapter, we explore the use of large language models for dementia detection from speech. We propose a method that leverages the capabilities of large language models to extract features from speech data and detect dementia. We evaluate our method on a

dataset of dementia patients and healthy individuals, and compare it with existing methods.

The remainder of this chapter is organized as follows: Section 3.2 reviews the related work in dementia detection from speech, including traditional methods and recent advancements using machine learning and deep learning techniques; Section 3.3 introduces the proposed method for dementia detection based on anomaly detection; Section 3.4 describes the experiment settings and results; Section 3.5 concludes the chapter and discusses future work.

3.2 Related Work

Early research detect dementia from speech by focusing on the content of the speech, semantic fluency, and comprehension of events [118]. Some researchers also explored content-agnostic methods to detect dementia, such as [90], who extracted paralinguistic features using an automatic speech recognition system.

Works in recent years mostly used machine learning and deep learning to detect dementia from speech, in both audio form and text form. Many researchers utilized the Pitt Corpus [115] dataset that is discussed in Chapter 2 as the training set and benchmark to develop and evaluate their models. [98, 119, 120] used multiple language models to fit the speech of dementia patients and healthy subjects separately. They used the perplexities of the language models when making predictions on speech samples to detect dementia. Their works experimented with different parts of language components: semantics, syntax, or both. Their experiments revealed that the speech of dementia is distinguishable by either syntax or semantics.

Besides using the language model perplexity to detect dementia, many works are conducted using deep learning models to automatically extract features from the speech data. [121] implemented a transformer-based model to automatically extract linguistic features from the datasets and detect dementia. [122] used CNN models for audio classification and transformer models to do classifications for text. [123] implemented an

auto-encoder architecture model to extract features from the spectrogram, with a multi-perceptron layer network as the classifier to detect dementia using the bottleneck features of the auto-encoder. [124] proposed a method in which they extracted 17 kinds of features, which represent syntactic, semantic, lexical, and demographic information of the speech samples collected in the DementiaBank Pitt Corpus dataset. The relevance of the 17 features was quantified by calculating a mutual information score that can represent the dependency between the extracted features and the MMSE score. They concluded that their methodology is superior to the methods based on syntactic information or the BERT method with only linguistic features. [125] fine-tuned the wav2vec [126] and word2vec [127] models on the dementia speech and text dataset to predict the labels of the samples as either dementia or healthy. They conducted experiments with different pre-processing methods on the text data. Their experiment results showed that with synonym-based data augmentation methods, their system can reach the best performance in dementia detection.

Except for using the linguistic features, the ADReSSo challenge [128] was introduced to encourage developments in dementia detection using the audio data only (however, using ASR to transcribe the speech was encouraged). [129] characterized the speech audio using x-vectors and prosody features. Besides that, they also used an ASR system to transcribe the speech data and fine-tune the BERT models to classify and assess the speakers. [130] developed an AI-based end-to-end system to detect dementia directly from the audio recordings and predict the severity at the same time. In the center of their system is a pre-trained data2vec model that is capable of processing speech, vision, and text. Ilias et al. [131] conducted research on AD by employing transformer-based models, with BERT achieving the highest accuracy of 87.50% in detecting dementia. They proposed an interpretable method using siamese networks, developed multi-task learning models for simultaneous dementia identification and severity assessment, and conducted a detailed linguistic analysis to identify significant language differences between AD and non-AD patients. Bouazizi et al. [132] conducted experiments to detect dementia from transcribed speech by focusing on changes in the subject's center of attention while describing the

cookie theft image. This method utilized an LSTM neural network for classification and demonstrated improved accuracy over conventional approaches by analyzing speech patterns and employing data augmentation techniques.

Recent years have seen a great leap in the development of large language models. With ChatGPT [133] being the leading product, large language models showed superior performance in many NLP areas without further training or fine-tuning. Many researches have been conducted using large language models in dementia detection. Li et al. [134] conducted experiments using a novel method that pairs a pre-trained GPT-2 model with a degraded version (GPT-D) to differentiate language produced by cognitively healthy individuals and those with AD. This approach demonstrated near state-of-the-art performance on the “Cookie Theft” description task, generalized well to spontaneous conversations, and revealed dementia-related linguistic anomalies in GPT-D’s generated text. Runde et al. [135] conducted experiments to enhance AD detection using NLP on spontaneous speech, leveraging audio enhancement techniques and AI-based transcription methods. This approach utilized advanced transcription services and GPT-based embeddings to improve the accuracy and efficiency of noninvasive AD detection.

In recent years, autoencoders have been widely used to learn representations of normal data and identify anomalies based on reconstruction errors [136]. Autoencoders are a class of neural networks designed for unsupervised representation learning, where the objective is to learn an efficient encoding of input data by minimizing the reconstruction error. The architecture typically consists of two symmetric components: an encoder that transforms the input into a lower-dimensional latent representation, and a decoder that reconstructs the input from this representation. By introducing a bottleneck in the network architecture, autoencoders are compelled to extract the most salient features of the data, facilitating dimensionality reduction and feature extraction [137]. Unlike linear techniques such as Principal Component Analysis (PCA), autoencoders are capable of capturing complex, nonlinear relationships through the use of deep architectures and nonlinear activation functions. Variants such as sparse autoencoders [138], denoising autoencoders [139], convolutional autoencoders [140], and variational autoencoders [141]

further extend the basic model to address specific learning scenarios, including noise robustness, spatial feature preservation, and generative modeling. Due to their versatility and adaptability, autoencoders have been widely applied in domains such as image processing, anomaly detection, signal denoising, and unsupervised pretraining of deep neural networks.

The related works discussed above have shown that dementia detection from speech is a promising area of research, with many approaches being proposed to detect dementia from speech data. However, despite the advancements in the general NLP field, it is challenging to apply the most recent NLP techniques to dementia detection because they usually require large amounts of data to train the models, which is not available in the dementia detection or other clinical tasks. Therefore, motivated by the limitations of current limitations in the field of dementia detection, we propose a new approach that leverages the most recent advancements in large language models and anomaly detection to detect dementia from speech data.

The contributions of this chapter are summarized as follows:

- We proposed a new architecture that leverages the most recent large language models and anomaly detection techniques, specifically designed for dementia detection, where the data is limited.
- We investigated the performance of multiple variants of autoencoders, including basic autoencoders, variational autoencoders and denoising autoencoders to learn the representations of the speech data.
- We explored and investigated the impact of different large language models, such as RoBERTa and LLaMA, on the performance of the proposed method.

3.3 Dementia Detection Based on Anomaly Detection

3.3.1 Language Model Embeddings

Language embeddings are dense, continuous vector representations of textual units such as words, phrases, sentences, or documents, designed to capture semantic and syntactic properties of language [142]. These representations are typically learned from large corpora using neural models that optimize context-based objectives, such as predicting surrounding words (e.g., Word2Vec [143]), maximizing likelihood of sequences (e.g., language models), or capturing masked token dependencies (e.g., BERT [92]). As a result, language embeddings encode contextual meaning, enabling downstream NLP tasks such as text classification, machine translation, and question answering to benefit from improved generalization and reduced data sparsity.

In this research, we utilized two pre-trained language models to extract the text embeddings, RoBERTa [144] and LLaMA [114], to extract dense vector embeddings from the speech samples in the Pitt Corpus dataset. These embeddings serve as compact representations of the linguistic features of the speech data, which are suitable for anomaly detection tasks using autoencoders. RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a variant of the BERT model introduced by Liu et al. [144] that aims to enhance the pretraining procedure and performance of BERT by making systematic modifications to its training methodology. RoBERTa retains the original architecture of BERT but improves upon its training efficiency and effectiveness through several key adjustments: training on significantly larger corpora, removing the Next Sentence Prediction (NSP) objective, utilizing larger mini-batches and learning rates, and dynamically changing the masking pattern applied to the input sequences during pretraining. Large Language Model Meta AI (LLaMA) [114] is a family of foundational language models developed by Meta AI that aim to provide high-performance natural language understanding and generation with relatively lower computational and data requirements compared to prior large-scale models. LLaMA models are trained on a diverse and extensive corpus, enabling them to capture a wide range of linguistic patterns and knowledge. They

are designed to be efficient and effective for various NLP tasks, including text classification, summarization, translation, and question answering. The LLaMA family includes multiple model sizes, allowing users to choose the appropriate model based on their computational resources and specific application requirements.

3.3.2 Autoencoders for Anomaly Detection

In this research, we employed autoencoders as the primary method for anomaly detection in dementia detection tasks.

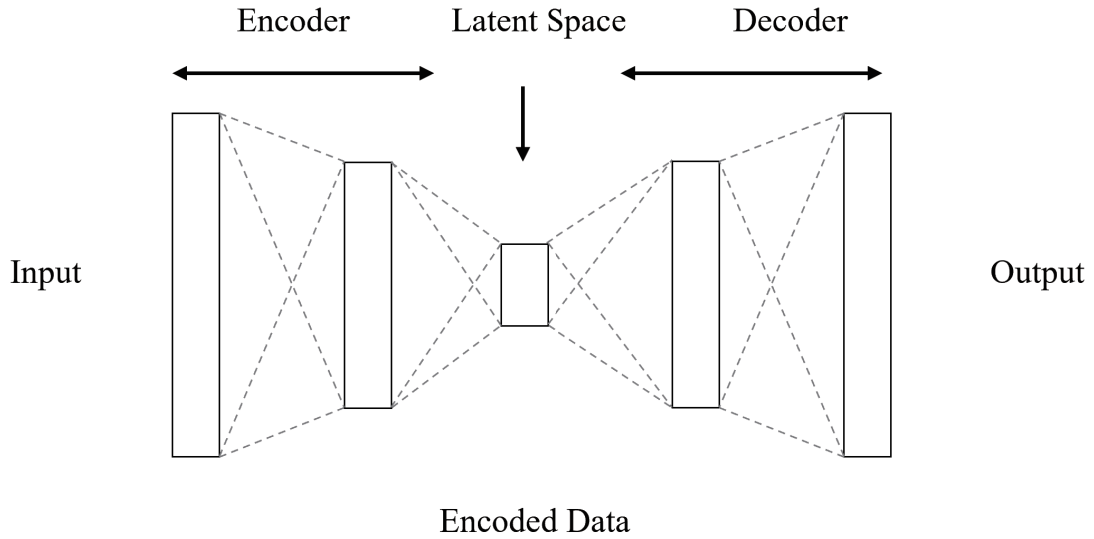


Figure 3.1. Architecture of the autoencoder used for anomaly detection in dementia detection tasks. The encoder maps the input embeddings to a lower-dimensional latent space, while the decoder reconstructs the original embeddings from this latent representation.

Figure 3.1 illustrates the basic architecture of the autoencoder used in this research. The autoencoder consists of two main components: the encoder and the decoder. The encoder maps the input embeddings to a lower-dimensional latent space, while the decoder reconstructs the original embeddings from this latent representation. The objective of the autoencoder is to minimize the reconstruction error between the input embeddings and their reconstructions, which allows the model to learn a compact representation of the data.

3.3.3 Dataset and Preprocessing

In this research, we conducted experiments on the Pitt Corpus [115], a widely used dataset for dementia detection, which contains speech samples from both healthy individuals and dementia patients, as introduced and explained in Chapter 2. The dataset consists of audio recordings of subjects describing the Cookie Theft picture, a standard task in dementia assessment. In this research, to leverage the capabilities of large language models, we used the pre-trained RoBERTa and LLaMA models to extract dense vector embeddings from the speech samples. The embeddings were extracted from the last hidden layer of the models, which captures the contextual information of the speech samples. Compared to the raw text data, the dense vector embeddings provide a more compact and informative representation of the speech samples, which is suitable for anomaly detection tasks using autoencoders. Most importantly, the embeddings are fixed-length vectors, which allows us to use them as input to the autoencoder without further preprocessing.

3.3.4 Anomaly Detection Approach

The proposed zero-shot approach for dementia detection is based on anomaly detection using autoencoders. The main idea is to train an autoencoder on the embeddings of healthy speech samples, allowing it to learn a compact representation of normal speech patterns. During inference, the autoencoder is used to reconstruct the embeddings of both healthy and dementia samples. The reconstruction error, measured as the mean squared error (MSE) between the input embeddings and their reconstructions, serves as an anomaly detection signal. As the autoencoder is trained only on healthy samples, it is expected to reconstruct these samples with low error. However, when presented with dementia samples, the reconstruction error is likely to be higher due to the differences in speech patterns associated with dementia, as shown in Fig. 3.2. Therefore, the reconstruction error can be used as a criterion for classification. The classification decision can be made by setting a threshold on the reconstruction error or using a regression or classification model to map the reconstruction error to a binary label (healthy or dementia).

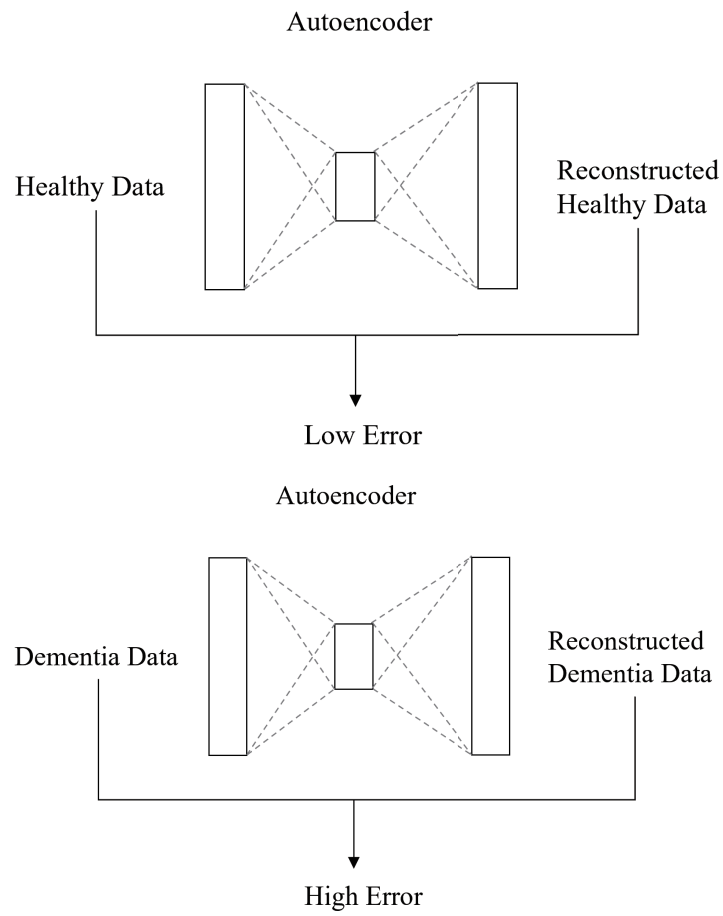


Figure 3.2. Overview of the anomaly detection approach for dementia detection. The autoencoder is trained on healthy samples, and the reconstruction error is used as an anomaly detection signal to classify dementia samples.

3.4 Experiment Settings and Results

3.4.1 Dementia Detection with Healthy Samples Only

Dataset and Preprocessing

Experiments were conducted on a preprocessed dataset derived from the Pitt Corpus, containing fixed-length embeddings extracted using a pre-trained RoBERTa model. Each sample is associated with a subject identifier, a binary label (0 = healthy control, 1 = dementia), and a dense vector embedding. These embeddings represent linguistic features and were stored as numerical arrays in a serialized format for efficient loading using pandas.

Model Architecture

A symmetric autoencoder was employed to learn a compact representation of healthy speech embeddings. The encoder consisted of two fully connected layers: the first mapped the input to 256 hidden units followed by a ReLU activation; the second projected to a lower-dimensional latent space with another ReLU activation. The decoder mirrored this structure to reconstruct the original embedding.

Training Procedure

The autoencoder was trained using only healthy control samples. The objective was to minimize the mean squared error (MSE) between input and reconstructed embeddings. Optimization was performed using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 64. Training was conducted for a maximum of 400 epochs per fold, with evaluation performed at each epoch.

Evaluation Metrics and Thresholding Strategy

Reconstruction loss was used as an anomaly detection signal. At each epoch, the model was evaluated on both healthy and dementia-labeled samples from the test fold. For each

sample, the total reconstruction loss (summed over all embedding dimensions) was computed. Although in the decision of classification, only 1-dimensional reconstruction loss is used, we still used a logistic regression model to map the reconstruction loss to a binary label for evaluation purposes.

Cross-Validation Protocol

A five-fold group cross-validation scheme was used, with grouping based on subject identifiers to prevent subject overlap between training and test sets. For each fold, only healthy samples from the training set were used to train the autoencoder. The test set for each fold included both healthy and dementia samples. This protocol was designed to control for subject-level variability and ensure reliable generalization performance.

Hyperparameter Tuning

The dimensionality of the latent representation was treated as a tunable hyperparameter. Experiments were repeated for latent dimensions in the set {8, 16, 32, 64, 128, 256, 512}. For each configuration, the best test accuracy (based on the optimal threshold) was recorded per fold. For each latent dimension, the model was trained for 200 epochs and evaluated on all five folds, and the average performance metrics were computed.

Experiment Results

Table 3.1, 3.2, and 3.3 summarize the performance metrics across different latent dimensions for the autoencoder trained on several embeddings.

The results in Table 3.1, 3.2, and 3.3 show the average accuracy, ROC AUC, and per-class precision, recall, and F1-score across different latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings. The results indicate that the performance of the autoencoder varies with the latent dimension size, with a notable peak in accuracy 0.7210 at a latent dimension of 16 for the RoBERTa embeddings, and 0.7410 at a latent dimension of 256 for the LLaMA 3.1 embeddings, and 0.7590 at a latent dimension of 256 for the LLaMA 3.2 embeddings.

Table 3.1. Average performance across latent dimensions: accuracy, ROC AUC, and class-specific precision, recall, and F1-score for the autoencoder trained on RoBERTa embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7174	0.7512	0.7086	0.6105	0.6226	0.7410	0.7495	0.7369
16	0.7210	0.7589	0.7446	0.5853	0.6161	0.7316	0.7747	0.7449
32	0.7137	0.7548	0.7997	0.5255	0.5529	0.7208	0.7996	0.7470
64	0.7193	0.7596	0.7436	0.5804	0.6094	0.7301	0.7746	0.7441
128	0.7120	0.7540	0.7137	0.5528	0.5847	0.7202	0.7837	0.7443
256	0.7084	0.7419	0.7576	0.4843	0.5326	0.7085	0.8303	0.7536
512	0.7156	0.7566	0.7384	0.5407	0.5634	0.7284	0.7911	0.7488

Table 3.2. Average performance across latent dimensions: accuracy, ROC AUC, and class-specific precision, recall, and F1-score for the autoencoder trained on LLaMA 3.1 7B embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7229	0.7486	0.7649	0.5603	0.6046	0.7234	0.7987	0.7532
16	0.7193	0.7546	0.7419	0.5734	0.6004	0.7310	0.7782	0.7458
32	0.7301	0.7659	0.7701	0.5651	0.6047	0.7325	0.8061	0.7612
64	0.7374	0.7630	0.7765	0.5815	0.6192	0.7416	0.8061	0.7657
128	0.7392	0.7664	0.7673	0.5915	0.6273	0.7443	0.8041	0.7669
256	0.7410	0.7715	0.7649	0.6015	0.6349	0.7471	0.8012	0.7675
512	0.7374	0.7685	0.7782	0.5953	0.6210	0.7496	0.7908	0.7600

Table 3.3. Average performance across latent dimensions: accuracy, ROC AUC, and class-specific precision, recall, and F1-score for the autoencoder trained on LLaMA 3.2 7B embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7536	0.7661	0.7995	0.6209	0.6677	0.7540	0.8125	0.7765
16	0.7518	0.7744	0.7805	0.6200	0.6695	0.7489	0.8148	0.7765
32	0.7554	0.7946	0.7868	0.6196	0.6659	0.7548	0.8171	0.7800
64	0.7536	0.7678	0.7776	0.6308	0.6733	0.7560	0.8072	0.7760
128	0.7536	0.7736	0.8065	0.6169	0.6604	0.7563	0.8117	0.7763
256	0.7590	0.7784	0.7970	0.6291	0.6690	0.7641	0.8124	0.7811
512	0.7536	0.7801	0.7902	0.6117	0.6616	0.7515	0.8208	0.7799

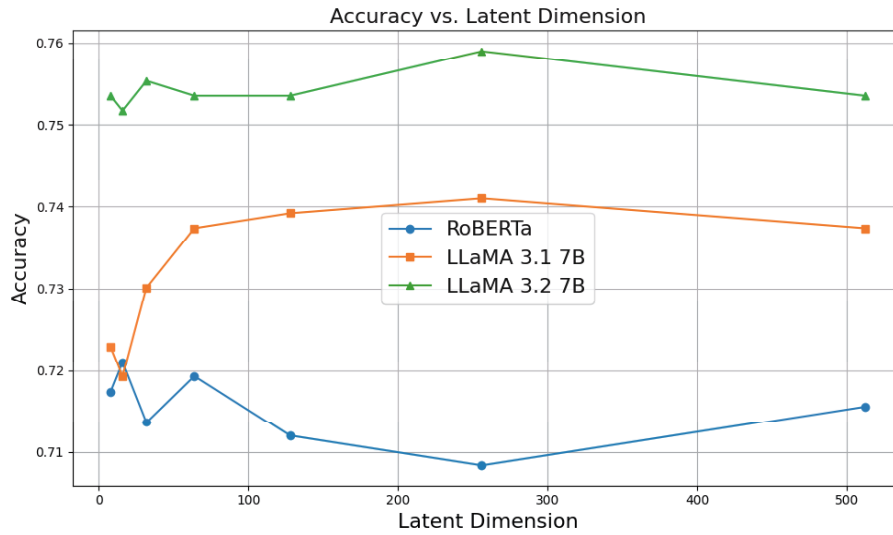


Figure 3.3. Average accuracy across latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings.

Figure 3.3 visualized the average accuracy across latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings. Despite the variability in performance across different latent dimensions, the overall trend suggests that a moderate selection of the latent space size is beneficial. Besides, the results show that when using LLaMA embeddings, the autoencoder generally achieves higher accuracy and F1 scores compared to when using RoBERTa embeddings, indicating that LLaMA embeddings may capture more relevant features for dementia detection. It is worth noting that the best performance was achieved with a latent dimension of 256 for LLaMA embeddings, which suggests that this dimensionality may be optimal for capturing the underlying structure of the data while maintaining sufficient discriminative power. Compared to the size of latent dimension settings, the embedding type has a more significant impact on the performance of the autoencoder, with LLaMA 3.2 embeddings consistently outperforming RoBERTa embeddings across all latent dimensions.

Figure 3.4 and 3.5 illustrates the reconstruction loss distributions for healthy and dementia samples before and after training, using a latent dimension of 64. The top plot shows the initial reconstruction losses before training, while the bottom plot shows the losses after 200 epochs of training. The separation between healthy and dementia samples becomes more pronounced after training, indicating that the autoencoder effectively

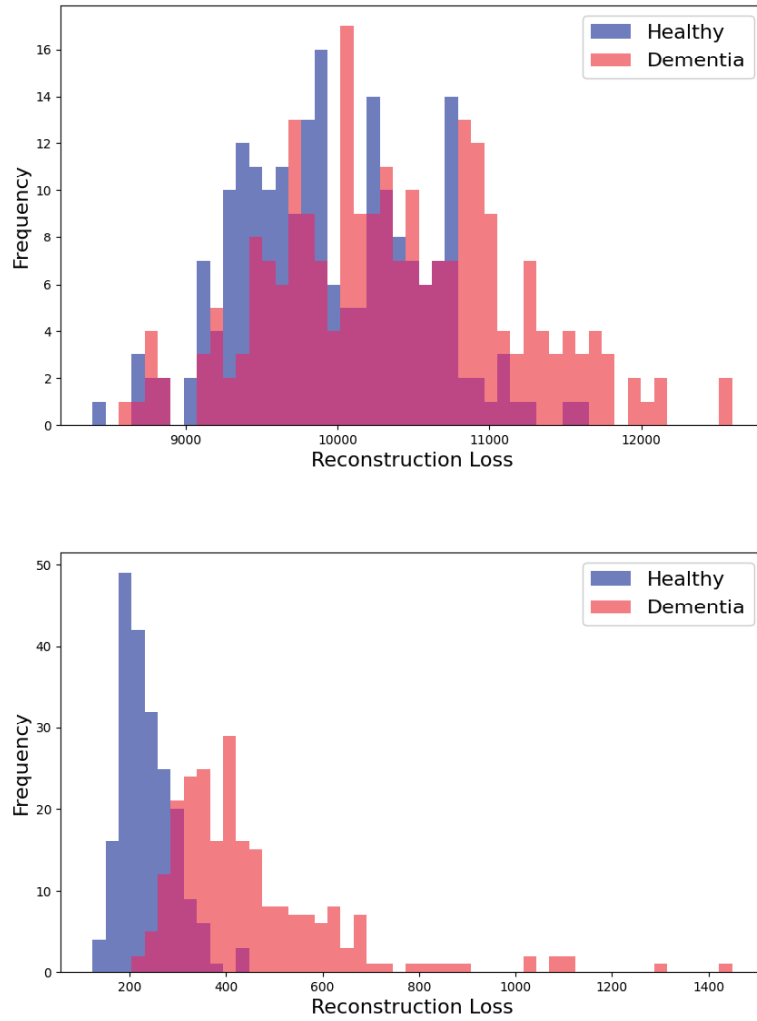


Figure 3.4. Reconstruction loss distributions for healthy and dementia samples in fold 2, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.

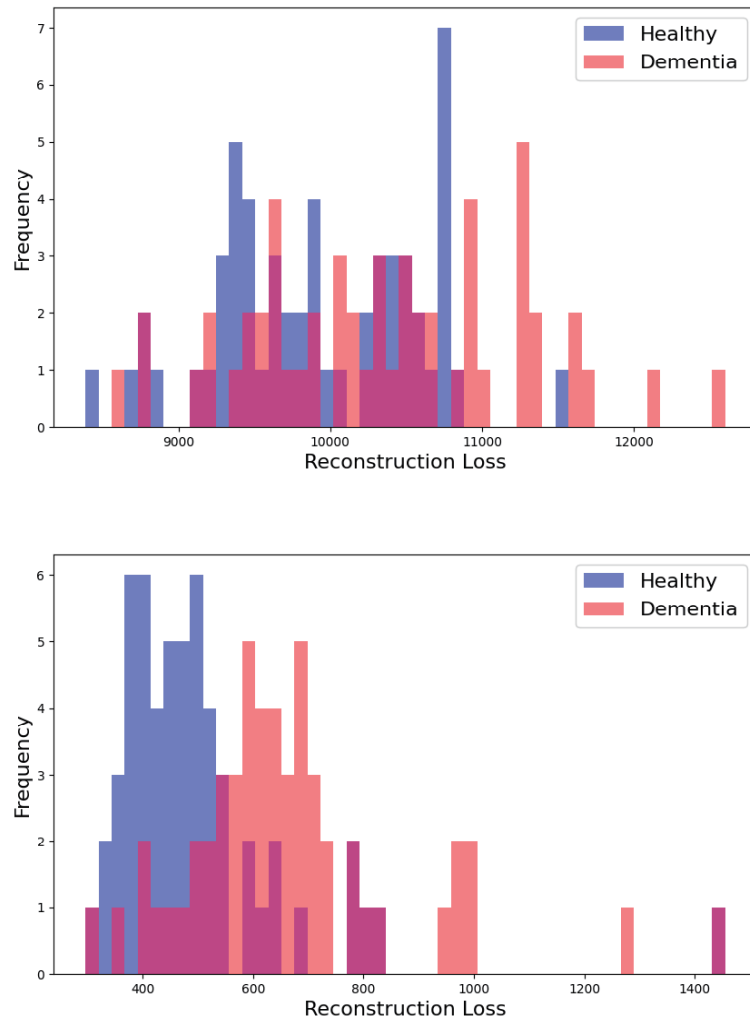


Figure 3.5. Reconstruction loss distributions for healthy and dementia samples in fold 4, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.

learned to reconstruct healthy embeddings while producing higher losses for dementia samples.

3.4.2 Dementia Detection with Healthy and Dementia Samples

In this section, the approach was extended to include both healthy and dementia samples in the training of the autoencoder. The goal was to evaluate whether the autoencoder could effectively learn to reconstruct embeddings from both classes and distinguish between them based on reconstruction loss.

Experiment Settings

The architecture of the autoencoder remained the same as in the previous experiment, with a symmetric structure consisting of an encoder and decoder. The encoder maps the input embeddings to a lower-dimensional latent space, while the decoder reconstructs the original embeddings from this latent representation. However, in this experiment, instead of one autoencoder trained only on healthy samples, two separate autoencoders were trained: one on healthy samples and another on dementia samples. The training procedure involved minimizing the mean squared error (MSE) between the input embeddings and their reconstructions for each class-specific autoencoder. The training was conducted using the same Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 64. The training was performed for a maximum of 400 epochs, with evaluation at each epoch to monitor the reconstruction loss. During evaluation, two losses (one for autoencoder trained on healthy samples and another for the one trained on dementia samples) were computed for each sample in the test set. The classification decision was made based on logistic regression to evaluate the performance for dementia detection.

Results

The results in Table 3.4, 3.5, and 3.6 show the average accuracy, ROC AUC, and per-class precision, recall, and F1-score across different latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings using two separate autoencoders for

Table 3.4. Cross-validation performance summary across latent dimensions when using RoBERTa embeddings, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7644	0.7879	0.7628	0.6844	0.7126	0.7726	0.7962	0.7814
16	0.7735	0.8072	0.7731	0.6939	0.7170	0.7908	0.7987	0.7890
32	0.7662	0.7972	0.7573	0.6932	0.7132	0.7831	0.7904	0.7816
64	0.7699	0.8055	0.7696	0.6906	0.7142	0.7838	0.7956	0.7854
128	0.7753	0.7923	0.7658	0.7110	0.7277	0.7930	0.7935	0.7890
256	0.7681	0.8088	0.7610	0.6882	0.7125	0.7806	0.7973	0.7857
512	0.7699	0.8069	0.7692	0.6865	0.7085	0.7877	0.7954	0.7862

Table 3.5. Cross-validation performance summary across latent dimensions when using LLaMA 3.1 embeddings, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7626	0.8015	0.7294	0.7061	0.7144	0.7827	0.7826	0.7811
16	0.7789	0.8229	0.7318	0.7558	0.7414	0.8140	0.7726	0.7911
32	0.7989	0.8220	0.7545	0.7737	0.7609	0.8351	0.7926	0.8110
64	0.8043	0.8476	0.7593	0.7792	0.7668	0.8393	0.8012	0.8175
128	0.8060	0.8603	0.7736	0.7522	0.7587	0.8278	0.8244	0.8242
256	0.8006	0.8406	0.7586	0.7693	0.7604	0.8336	0.7996	0.8137
512	0.8079	0.8537	0.7689	0.7708	0.7675	0.8349	0.8144	0.8232

Table 3.6. Cross-validation performance summary across latent dimensions when using LLaMA 3.2 embeddings, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7735	0.8109	0.7688	0.6953	0.7212	0.7810	0.8087	0.7919
16	0.7790	0.8108	0.7547	0.7374	0.7398	0.8025	0.7850	0.7902
32	0.7916	0.8299	0.7688	0.7315	0.7462	0.8062	0.8156	0.8091
64	0.7988	0.8402	0.7687	0.7673	0.7649	0.8232	0.8007	0.8099
128	0.8006	0.8303	0.7754	0.7605	0.7644	0.8194	0.8111	0.8138
256	0.7988	0.8298	0.7707	0.7584	0.7613	0.8198	0.8081	0.8119
512	0.7970	0.8337	0.7750	0.7462	0.7566	0.8127	0.8146	0.8118

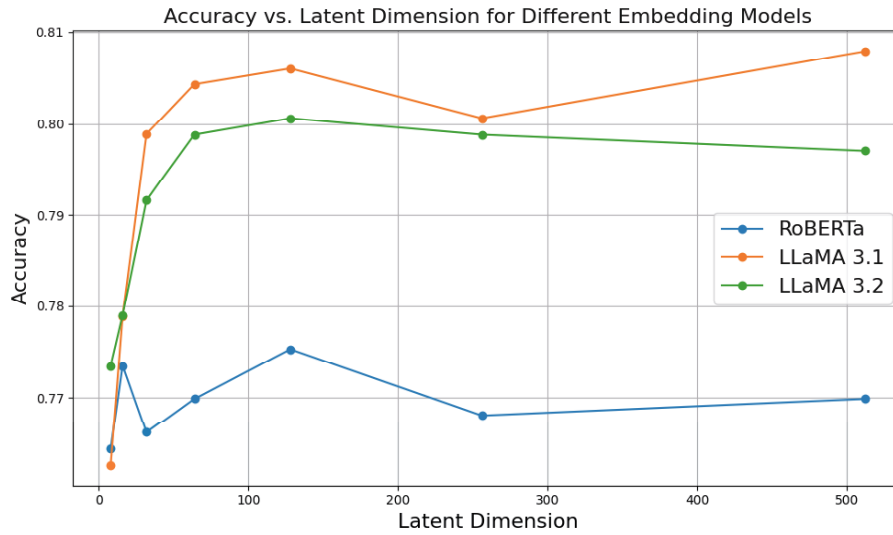


Figure 3.6. Average accuracy across latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings, using two separate autoencoders for healthy and dementia samples.

healthy and dementia samples. The results indicate that the performance of the autoencoder varies with the latent dimension size, with a notable peak in accuracy 0.7753 at a latent dimension of 128 for the RoBERTa embeddings, 0.8079 at a latent dimension of 512 for the LLaMA 3.1 embeddings, and 0.8006 at a latent dimension of 128 for the LLaMA 3.2 embeddings. Figure 3.6 illustrates the average accuracy across latent dimensions for the autoencoder trained on RoBERTa and LLaMA embeddings using two separate autoencoders for healthy and dementia samples. The results show that the performance of the autoencoder is generally better when using LLaMA embeddings compared to RoBERTa embeddings, with the best performance achieved with a latent dimension of 512 for LLaMA 3.1 embeddings and 128 for LLaMA 3.2 embeddings. The figure also indicate that the impact of the latent dimension size on the performance of the autoencoder is not as pronounced when the size is over 64, suggesting that the autoencoder can effectively learn to reconstruct embeddings from both classes with a relatively small latent dimension size.

Distribution of Reconstruction Losses Analysis

Figures 3.7 and 3.8 illustrate the reconstruction loss distributions for healthy and dementia samples before and after training, using a latent dimension of 64. The top plot shows

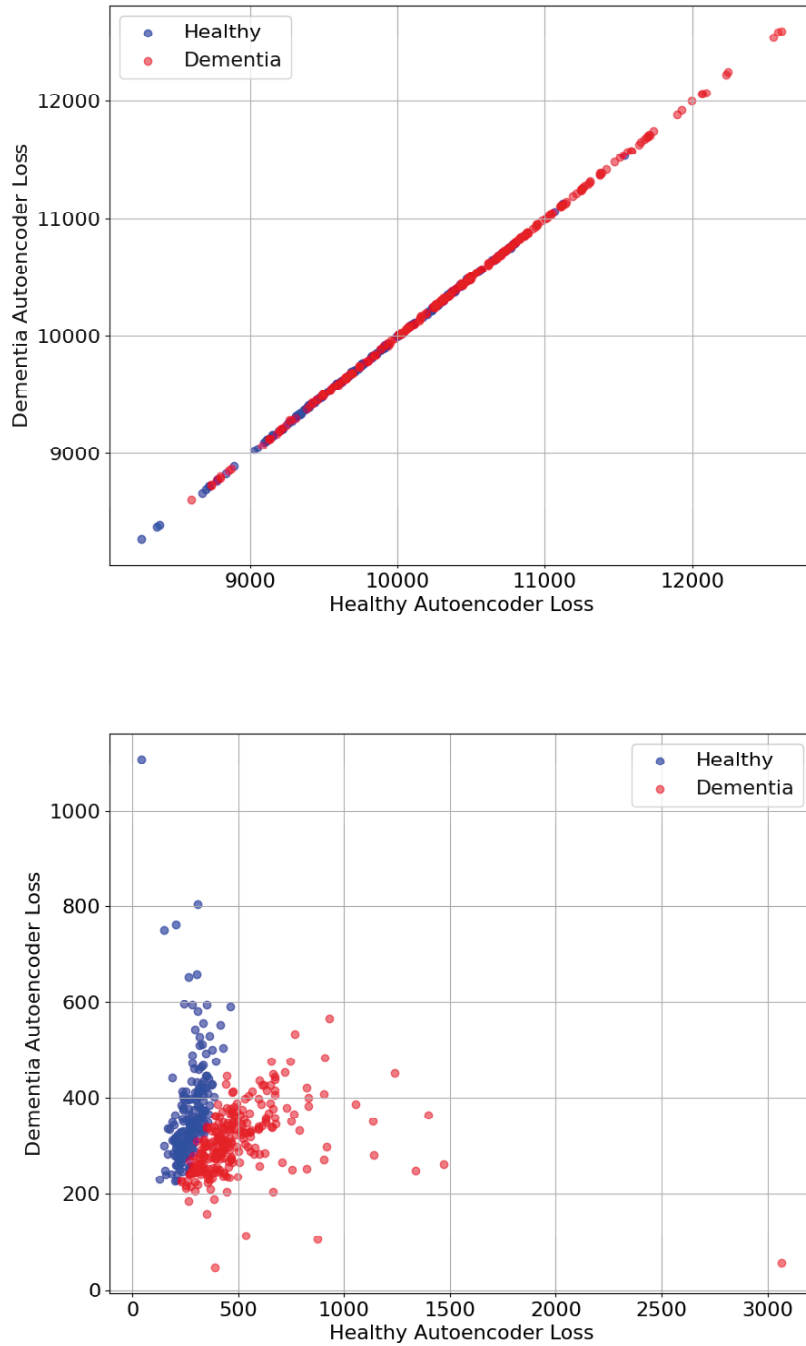


Figure 3.7. Reconstruction loss distributions for healthy and dementia samples using two separate autoencoders, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.

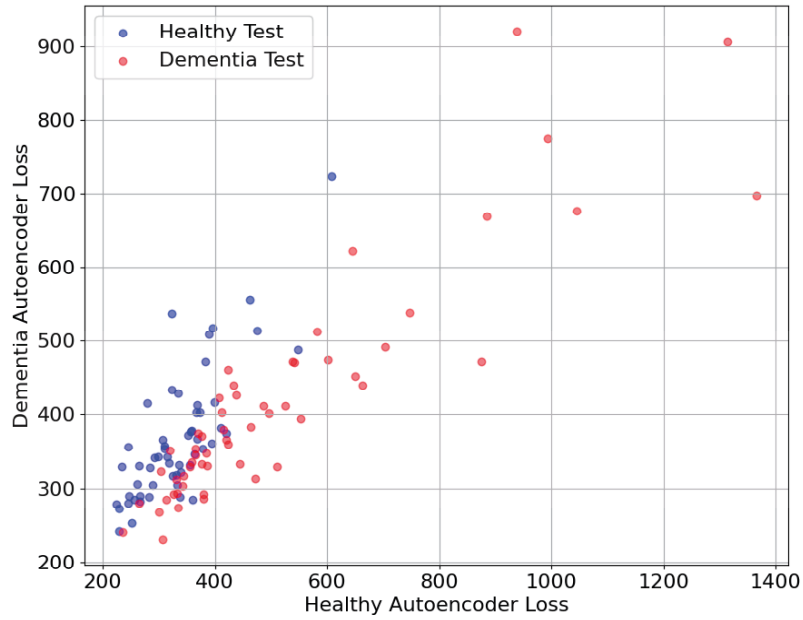
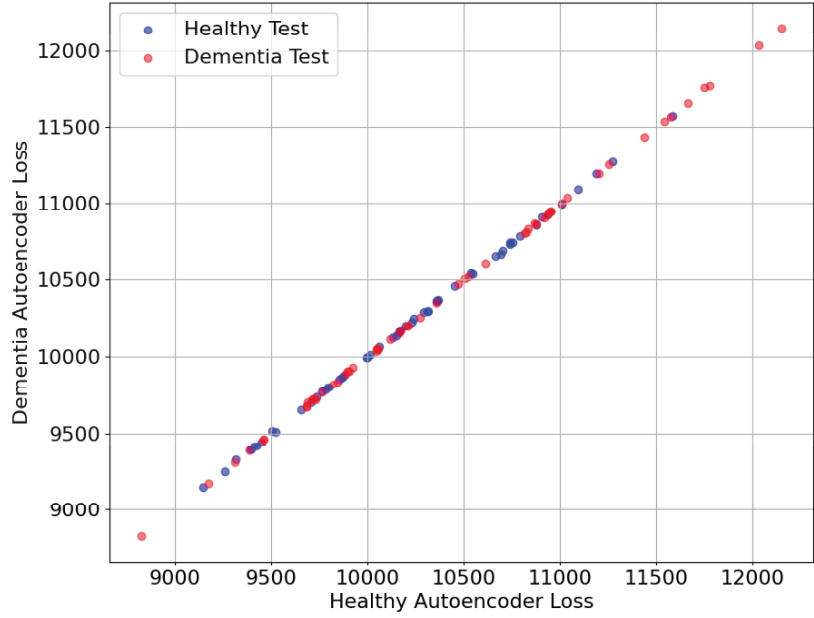
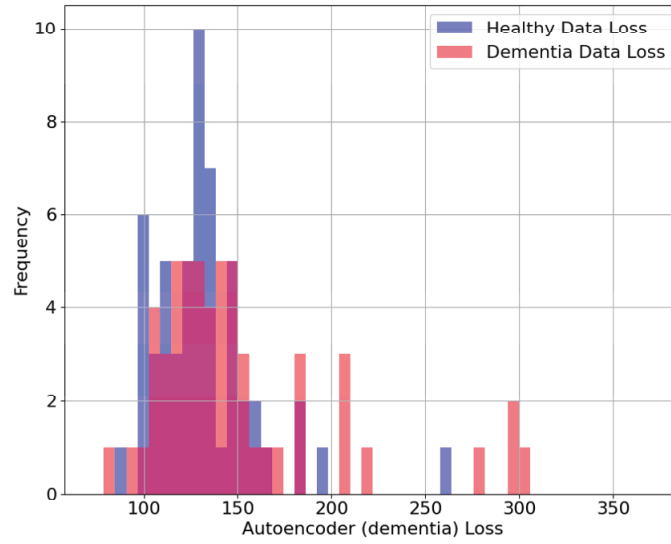
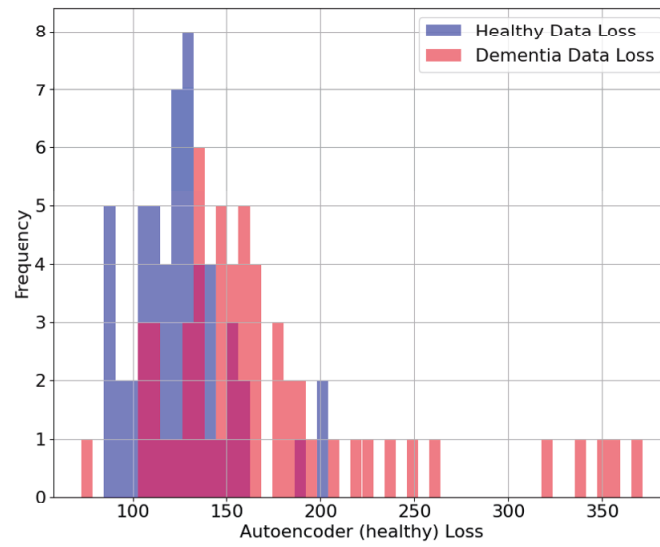


Figure 3.8. Reconstruction loss distributions for healthy and dementia samples using two separate autoencoders, with latent dimension 64. The top plot shows the initial reconstruction losses, while the bottom plot shows losses after training.



(a)



(b)

Figure 3.9. Histogram of reconstruction losses for dementia and healthy samples using two separate autoencoders, with latent dimension 256. (a) shows the distribution for reconstruction errors under the autoencoder trained on dementia data, while (b) shows the distribution for reconstruction errors under the autoencoder trained on healthy data. The x-axis represents the reconstruction loss, and the y-axis represents the number of samples.

the initial reconstruction losses before training, while the bottom plot shows the losses after 170 epochs of training. Before training, the loss distributions for healthy and dementia samples overlap significantly, making it challenging to distinguish between the two classes. The separation between healthy and dementia samples becomes more pronounced after training, indicating that the autoencoder effectively learned to reconstruct embeddings from both classes while producing higher losses for dementia samples. Figure 3.9 shows the separate histograms of reconstruction losses for dementia and healthy samples using two separate autoencoders, with a latent dimension of 256. The results show that autoencoders trained on healthy samples have better ability to distinguish between healthy and dementia samples, while the autoencoders trained on dementia samples, although still effective, have a higher overlap in reconstruction losses between the two classes. This suggests that the autoencoder trained on healthy samples is more effective in capturing the characteristics of healthy embeddings.

3.4.3 Autoencoder Variants

In this experiment, we compared the performance of different autoencoder variants, including baseline autoencoders, deeper autoencoders, denoising autoencoders, and variational autoencoders. Each variant was evaluated across various latent dimensions to determine their effectiveness in detecting dementia-related anomalies.

Experiment Settings

The experiment was conducted using the same dataset and preprocessing steps as in the previous experiments, with fixed-length embeddings extracted using a pre-trained RoBERTa model. The autoencoder variants were implemented as follows:

- **Baseline Autoencoder:** A symmetric autoencoder with two fully connected layers in the encoder and decoder, as described in the previous sections.
- **Deeper Autoencoder:** An autoencoder with an additional hidden layer in both the encoder and decoder, resulting in a deeper architecture.

- **Denoising Autoencoder:** An autoencoder that adds Gaussian noise to the input embeddings during training to encourage robustness against noise.
- **Variational Autoencoder (VAE):** A probabilistic autoencoder that learns a distribution over the latent space, allowing for sampling and generation of new embeddings.

Results

Table 3.7. Performance summary across latent dimensions when using deep autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7752	0.8103	0.7507	0.7267	0.7348	0.7925	0.7895	0.7887
16	0.7770	0.8160	0.7429	0.7383	0.7382	0.8003	0.7855	0.7908
32	0.7771	0.8024	0.7962	0.6831	0.7122	0.7826	0.8208	0.7971
64	0.7879	0.8227	0.7611	0.7420	0.7484	0.8063	0.8008	0.8021
128	0.7735	0.8145	0.7476	0.7202	0.7316	0.7883	0.7951	0.7904
256	0.7843	0.8116	0.7649	0.7345	0.7473	0.7960	0.8040	0.7989
512	0.7807	0.8207	0.7680	0.7136	0.7370	0.7863	0.8145	0.7990

Table 3.8. Performance summary across latent dimensions when using denoising (noise level 0.1) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7843	0.8120	0.7601	0.7380	0.7467	0.8004	0.8015	0.7993
16	0.7771	0.8148	0.7630	0.7182	0.7335	0.7897	0.7987	0.7919
32	0.7934	0.8243	0.7770	0.7359	0.7518	0.8041	0.8176	0.8095
64	0.8042	0.8232	0.7795	0.7678	0.7677	0.8298	0.8053	0.8142
128	0.7987	0.8387	0.7774	0.7395	0.7514	0.8186	0.8170	0.8144
256	0.7988	0.8261	0.7757	0.7539	0.7600	0.8178	0.8099	0.8117
512	0.8060	0.8439	0.7775	0.7606	0.7658	0.8262	0.8191	0.8211

Table 3.7, 3.8, 3.9, 3.10, 3.11, and 3.12 present the average accuracy, ROC AUC, and per-class precision, recall, and F1-score across different latent dimensions for the various

Table 3.9. Performance summary across latent dimensions when using denoising (noise level 0.2) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7771	0.8038	0.7722	0.6941	0.7278	0.7780	0.8259	0.7991
16	0.7790	0.8171	0.7611	0.7227	0.7352	0.7954	0.7949	0.7926
32	0.7861	0.8234	0.7578	0.7498	0.7487	0.8107	0.7893	0.7973
64	0.7916	0.8243	0.7605	0.7567	0.7556	0.8138	0.7983	0.8045
128	0.7988	0.8274	0.7833	0.7349	0.7557	0.8068	0.8293	0.8168
256	0.7987	0.8255	0.7925	0.7409	0.7545	0.8127	0.8185	0.8124
512	0.8024	0.8372	0.7752	0.7579	0.7639	0.8210	0.8154	0.8169

Table 3.10. Performance summary across latent dimensions when using denoising (noise level 0.3) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7735	0.7987	0.7531	0.7045	0.7261	0.7820	0.8107	0.7947
16	0.7790	0.8147	0.7588	0.7385	0.7405	0.8046	0.7794	0.7874
32	0.7952	0.8204	0.7669	0.7548	0.7575	0.8161	0.8047	0.8085
64	0.8043	0.8318	0.7863	0.7607	0.7682	0.8238	0.8119	0.8152
128	0.8042	0.8518	0.7764	0.7700	0.7694	0.8289	0.8072	0.8150
256	0.7969	0.8402	0.7887	0.7296	0.7513	0.8051	0.8266	0.8136
512	0.8079	0.8415	0.7733	0.7749	0.7696	0.8392	0.8081	0.8201

Table 3.11. Performance summary across latent dimensions when using denoising (noise level 0.4) autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7752	0.8063	0.7619	0.7081	0.7305	0.7823	0.8077	0.7925
16	0.7807	0.8145	0.7625	0.7256	0.7394	0.7946	0.7992	0.7949
32	0.7916	0.8342	0.7706	0.7389	0.7506	0.8059	0.8113	0.8071
64	0.7988	0.8277	0.7829	0.7533	0.7611	0.8177	0.8085	0.8101
128	0.7970	0.8272	0.7821	0.7427	0.7549	0.8141	0.8119	0.8100
256	0.7988	0.8367	0.7839	0.7420	0.7538	0.8172	0.8157	0.8128
512	0.8006	0.8388	0.7711	0.7598	0.7616	0.8247	0.8079	0.8137

Table 3.12. Performance summary across latent dimensions when using variational autoencoders, reporting accuracy, ROC AUC, and class-specific precision, recall, and F1-score.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7844	0.8149	0.7683	0.7121	0.7355	0.7924	0.8187	0.8042
16	0.7825	0.8170	0.7609	0.7407	0.7487	0.7961	0.7961	0.7953
32	0.7826	0.8159	0.7598	0.7171	0.7354	0.7936	0.8131	0.8024
64	0.7862	0.8172	0.7722	0.7123	0.7357	0.7967	0.8172	0.8050
128	0.7789	0.8136	0.7686	0.6990	0.7298	0.7808	0.8222	0.8004
256	0.7934	0.8175	0.7705	0.7455	0.7544	0.8106	0.8071	0.8073
512	0.7916	0.8266	0.7771	0.7327	0.7498	0.8038	0.8127	0.8064

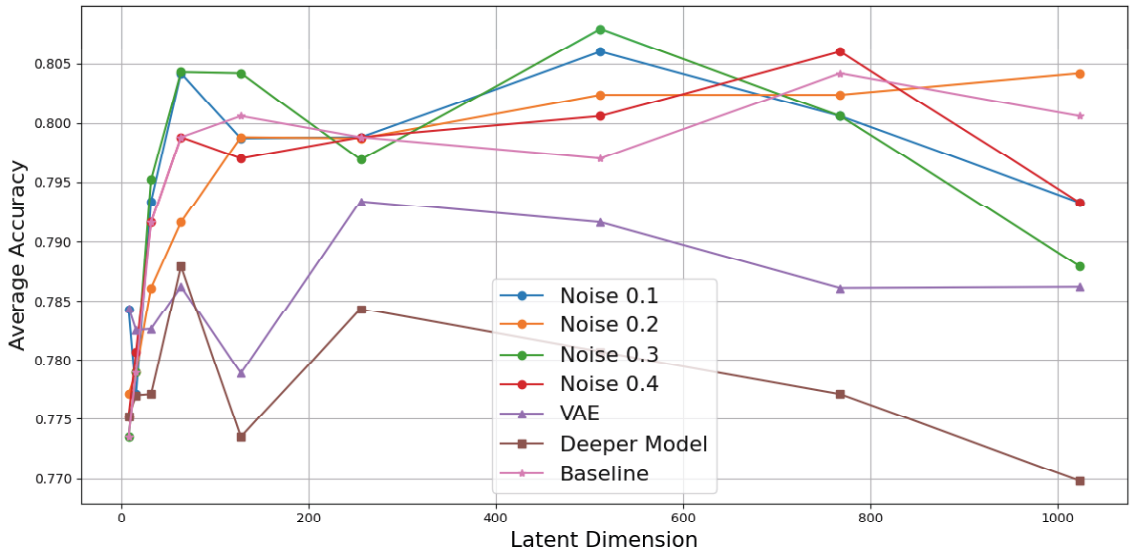


Figure 3.10. Average accuracy across latent dimensions for different autoencoder variants, including baseline, deeper autoencoders, denoising autoencoders, and variational autoencoders.

autoencoder variants trained on text embeddings of LLaMA3.2. Figure 3.10 illustrates the average accuracy across latent dimensions for the different autoencoder variants. The results show that denoising autoencoders with a noise level of 0.3 achieved the highest average accuracy of 0.8079.

Table 3.13. Classification performance across latent dimensions using variational autoencoders trained on LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7898	0.8139	0.7833	0.7157	0.7427	0.7956	0.8234	0.8075
16	0.7862	0.8048	0.7759	0.7083	0.7317	0.7986	0.8159	0.8038
32	0.7880	0.8077	0.7538	0.7504	0.7478	0.8132	0.7962	0.8026
64	0.7916	0.8081	0.7715	0.7306	0.7485	0.8015	0.8196	0.8098
128	0.7862	0.8109	0.7583	0.7342	0.7443	0.8017	0.8063	0.8033
256	0.7898	0.8156	0.7716	0.7312	0.7478	0.8004	0.8154	0.8067
512	0.7953	0.8090	0.7813	0.7317	0.7513	0.8057	0.8213	0.8118
768	0.7844	0.7970	0.7712	0.7115	0.7361	0.7917	0.8179	0.8033
1024	0.7898	0.8108	0.7678	0.7363	0.7480	0.8058	0.8072	0.8046

Table 3.14. Classification performance across latent dimensions using deep autoencoders trained on LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7971	0.8185	0.7647	0.7509	0.7524	0.8244	0.8046	0.8113
16	0.7970	0.8331	0.7697	0.7463	0.7531	0.8179	0.8118	0.8124
32	0.8025	0.8215	0.7664	0.7620	0.7609	0.8292	0.8098	0.8173
64	0.7953	0.8206	0.7703	0.7402	0.7496	0.8156	0.8114	0.8108
128	0.8115	0.8375	0.7854	0.7596	0.7677	0.8327	0.8254	0.8269
256	0.7952	0.8277	0.7653	0.7432	0.7495	0.8181	0.8082	0.8109
512	0.7971	0.8183	0.7584	0.7538	0.7539	0.8224	0.8083	0.8138
768	0.7989	0.8332	0.7701	0.7566	0.7600	0.8203	0.8087	0.8127
1024	0.7934	0.8238	0.7591	0.7558	0.7544	0.8189	0.7984	0.8068

Table 3.13, 3.14, 3.15, 3.16, 3.17, 3.18, and 3.19 present the average accuracy, ROC AUC, and per-class precision, recall, and F1-score across different latent dimensions for the various autoencoder variants trained on text embeddings of LLaMA3.1.

Table 3.15. Classification performance across latent dimensions using denoising autoencoders (noise level 0.1) trained on LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7663	0.7878	0.7333	0.6992	0.7012	0.7984	0.7801	0.7847
16	0.7953	0.8264	0.7718	0.7411	0.7513	0.8143	0.8105	0.8103
32	0.8025	0.8386	0.7692	0.7607	0.7597	0.8308	0.8085	0.8167
64	0.7952	0.8456	0.7574	0.7454	0.7473	0.8226	0.8068	0.8121
128	0.8079	0.8527	0.7634	0.7829	0.7708	0.8417	0.8042	0.8206
256	0.8024	0.8398	0.7674	0.7609	0.7605	0.8298	0.8103	0.8166
512	0.8151	0.8432	0.7777	0.7918	0.7828	0.8436	0.8140	0.8264
768	0.8024	0.8372	0.7691	0.7572	0.7590	0.8271	0.8126	0.8176
1024	0.8042	0.8288	0.7664	0.7744	0.7667	0.8358	0.8014	0.8158

Table 3.16. Classification performance across latent dimensions using denoising autoencoders (noise level 0.2) trained on LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 Precision	Class 0 Recall	Class 0 F1	Class 1 Precision	Class 1 Recall	Class 1 F1
8	0.7645	0.7811	0.7304	0.7079	0.7033	0.8008	0.7683	0.7790
16	0.7880	0.8183	0.7462	0.7678	0.7550	0.8194	0.7820	0.7989
32	0.8043	0.8234	0.7734	0.7529	0.7593	0.8266	0.8200	0.8212
64	0.8043	0.8360	0.7687	0.7707	0.7657	0.8336	0.8055	0.8167
128	0.8096	0.8564	0.7786	0.7616	0.7671	0.8305	0.8241	0.8254
256	0.8078	0.8444	0.7667	0.7715	0.7651	0.8401	0.8105	0.8221
512	0.8078	0.8454	0.7645	0.7899	0.7747	0.8429	0.8003	0.8184
768	0.8114	0.8510	0.7684	0.7792	0.7713	0.8416	0.8150	0.8266
1024	0.8024	0.8302	0.7617	0.7638	0.7593	0.8329	0.8070	0.8175

Table 3.17. Classification performance across latent dimensions using denoising autoencoders (noise level 0.3) trained on LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 Precision	Class 0 Recall	Class 0 F1	Class 1 Precision	Class 1 Recall	Class 1 F1
8	0.7680	0.8072	0.7355	0.7205	0.7263	0.7861	0.7866	0.7854
16	0.8007	0.8377	0.7720	0.7683	0.7630	0.8336	0.7954	0.8098
32	0.8151	0.8294	0.7855	0.7694	0.7732	0.8380	0.8263	0.8300
64	0.8006	0.8350	0.7586	0.7727	0.7633	0.8313	0.8005	0.8138
128	0.7970	0.8406	0.7548	0.7576	0.7523	0.8296	0.7996	0.8118
256	0.7970	0.8460	0.7543	0.7615	0.7555	0.8273	0.8007	0.8122
512	0.8115	0.8584	0.7702	0.7918	0.7791	0.8427	0.8077	0.8229
768	0.8006	0.8163	0.7889	0.7171	0.7369	0.8207	0.8290	0.8196
1024	0.8097	0.8428	0.7746	0.7678	0.7664	0.8389	0.8161	0.8243

Table 3.18. Classification performance across latent dimensions using denoising autoencoders (noise level 0.4) trained on LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 Precision	Class 0 Recall	Class 0 F1	Class 1 Precision	Class 1 Recall	Class 1 F1
8	0.7772	0.8190	0.7386	0.7278	0.7296	0.8028	0.7894	0.7941
16	0.7880	0.8173	0.7613	0.7288	0.7304	0.8196	0.7988	0.8036
32	0.8024	0.8355	0.7556	0.7958	0.7725	0.8435	0.7858	0.8104
64	0.8079	0.8490	0.7718	0.7685	0.7660	0.8387	0.8133	0.8223
128	0.8096	0.8508	0.7666	0.7794	0.7714	0.8394	0.8122	0.8243
256	0.8097	0.8417	0.7813	0.7601	0.7658	0.8333	0.8233	0.8244
512	0.8061	0.8491	0.7668	0.7695	0.7672	0.8299	0.8161	0.8224
768	0.8042	0.8388	0.7716	0.7649	0.7660	0.8266	0.8138	0.8185
1024	0.8006	0.8348	0.7593	0.7685	0.7622	0.8290	0.8050	0.8153

Table 3.19. Performance summary across latent dimensions using basic autoencoders with LLaMA3.1 embeddings.

Latent Dim	Accuracy	ROC AUC	Class 0 (healthy)			Class 1 (dementia)		
			Precision	Recall	F1	Precision	Recall	F1
8	0.7626	0.8015	0.7294	0.7061	0.7144	0.7827	0.7826	0.7811
16	0.7789	0.8229	0.7318	0.7558	0.7414	0.8140	0.7726	0.7911
32	0.7989	0.8220	0.7545	0.7737	0.7609	0.8351	0.7926	0.8110
64	0.8043	0.8476	0.7593	0.7792	0.7668	0.8393	0.8012	0.8175
128	0.8060	0.8603	0.7736	0.7522	0.7587	0.8278	0.8244	0.8242
256	0.8006	0.8406	0.7586	0.7693	0.7604	0.8336	0.7996	0.8137
512	0.8079	0.8537	0.7689	0.7708	0.7675	0.8349	0.8144	0.8232
768	0.7988	0.8398	0.7511	0.7715	0.7574	0.8376	0.7931	0.8118
1024	0.8079	0.8439	0.7786	0.7607	0.7643	0.8329	0.8191	0.8226

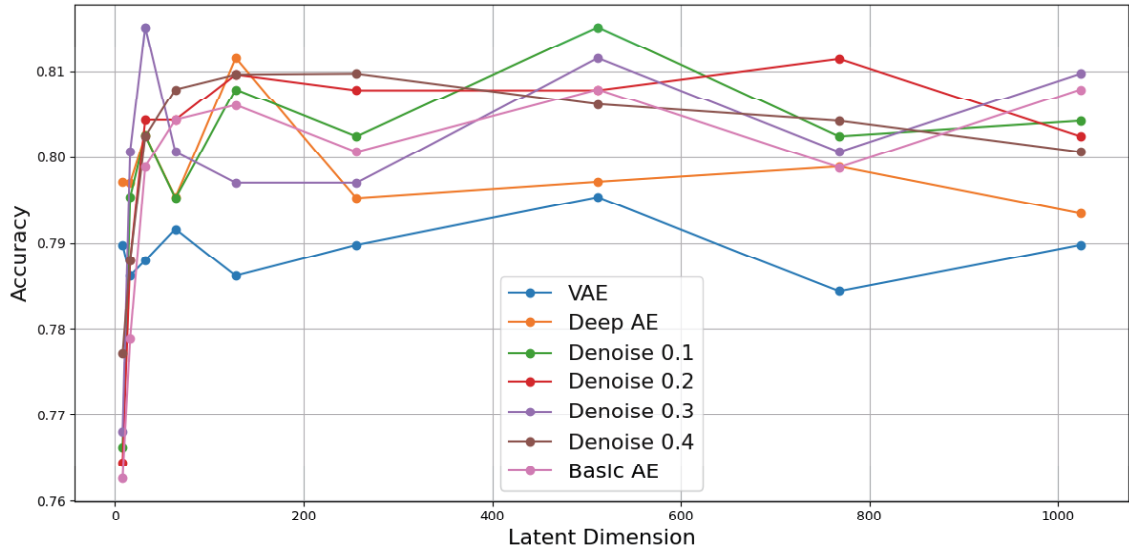


Figure 3.11. Average accuracy across latent dimensions for different autoencoder variants trained on LLaMA3.1 embeddings, including baseline, deeper autoencoders, denoising autoencoders, and variational autoencoders.

Figure 3.11 illustrates the average accuracy across latent dimensions for the different autoencoder variants trained on LLaMA3.1 embeddings. The results show that denoising autoencoders with a noise level of 0.3 and 0.1 achieved the highest average accuracy of 0.8151.

However, increasing the complexity of the autoencoder architecture, such as using deeper autoencoders or variational autoencoders, decreased the average accuracy to a level even lower than the baseline autoencoder. This suggests that while more complex architectures can capture intricate patterns in the data, they may also lead to overfitting, especially when the dataset is not sufficiently large or diverse. The denoising autoencoder achieved the best performance, indicating that adding noise during training can help the model generalize better to unseen data by adding more regulations to the training process.

Table 3.20 compares the accuracy of our method with previous studies on dementia-related data. Our approach demonstrates competitive performance, particularly when leveraging the strengths of large language model embeddings and advanced autoencoder architectures. Our method adapted the use of large language model embeddings, such as RoBERTa and LLaMA, which effectively capture the nuances of dementia-related data to the context of dementia detection, where the dataset is extremely scarce and the number of samples is limited. By employing autoencoders, we solved the problem of limited

Methods	Accuracy
Li et al. [134]	76%
Runde et al. [135]	82%
Ours (zero-shot)	76%
Ours (binary classification)	82%
Ours (by averaging each subject)	89%

Table 3.20. Comparison of accuracy across different CI detection methods.

data by unsupervised learning, allowing the model to learn meaningful representations of the data without requiring extensive labeled datasets. Based on this model, we achieved zero-shot classification with a high accuracy of 76% and a binary classification accuracy of 82%. Furthermore, by averaging the predictions across samples from the same subject, the accuracy can be improved to 89%, demonstrating the effectiveness of our approach in leveraging the strengths of large language model embeddings and advanced autoencoder architectures for dementia detection.

3.4.4 Error Analysis

Fig. 3.12 shows the ROC curve for the best-performing autoencoder variant (denoising autoencoder with noise level 0.3, final decisions were made by averaging across the samples of the same subject) trained on LLaMA 3.1 embeddings, achieving an AUC of 0.8913. The precision is 0.9512, sensitivity (recall) is 0.9017, specificity is 0.8806, F1 score is 0.9258, and accuracy is 0.8958 with a p-value of 9.342×10^{-28} . This indicates good performance in distinguishing between healthy and dementia-related samples.

Figures 3.13 and 3.14 illustrate the relationship between reconstruction error and MMSE scores. The first figure shows the distribution of reconstruction errors for correctly classified (blue) and misclassified (red) samples, while the second figure displays the misclassification rate across different MMSE scores. The analysis reveals that samples with MMSE scores near the boundary (around 27) exhibit higher misclassification rates, indicating that the model struggles to differentiate between healthy and dementia-related samples in this range. In practice, subjects with MMSE score higher than 27 are often

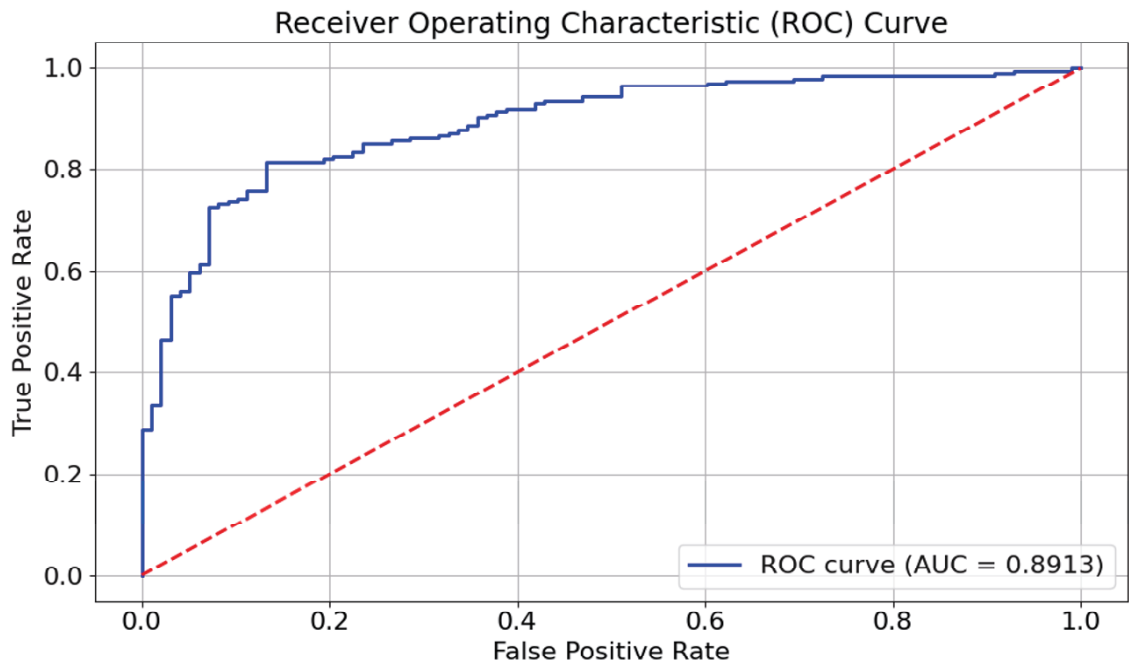


Figure 3.12. ROC curve for the best-performing autoencoder variant (denoising autoencoder with noise level 0.3) trained on LLaMA 3.1 embeddings. The area under the curve (AUC) is 0.8913, indicating good performance in distinguishing between healthy and dementia-related samples.

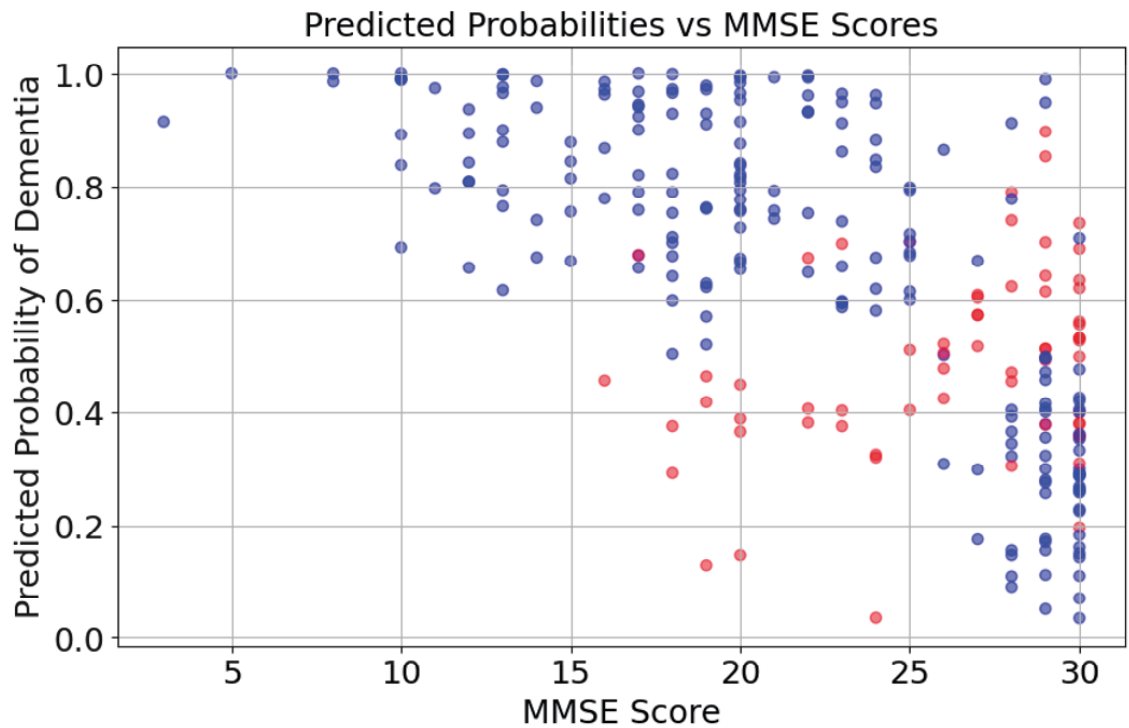


Figure 3.13. Difference in reconstruction error by MMSE score. Blue points represent correctly classified samples; red points are misclassified.

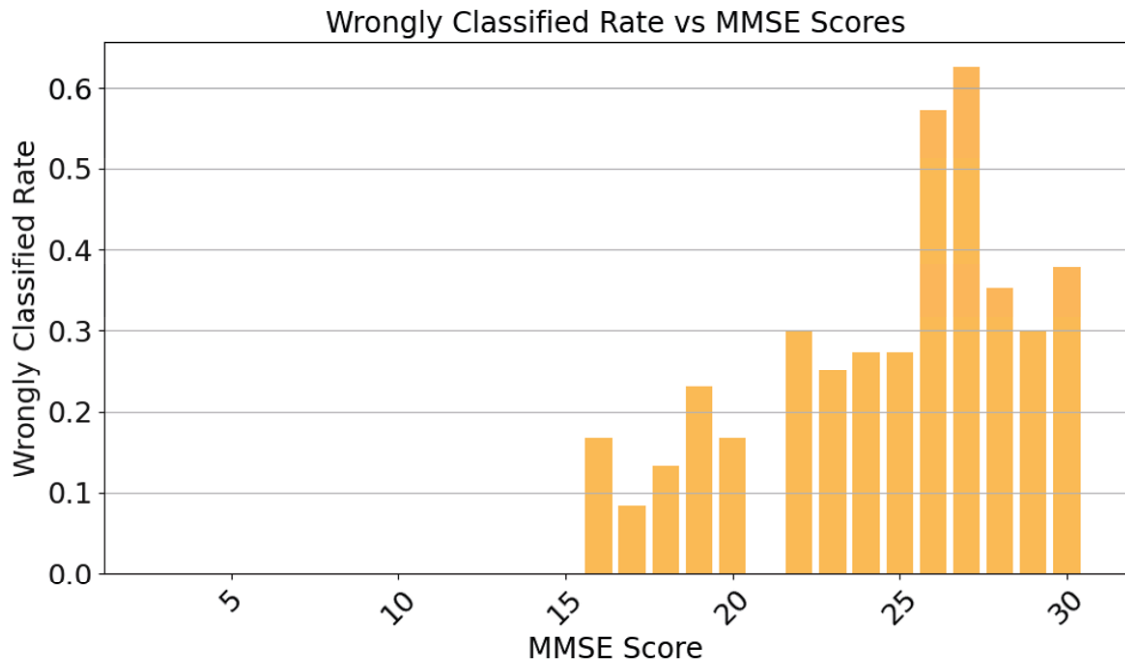


Figure 3.14. Misclassification rate by MMSE score. Patients with MMSE scores near the boundary (MMSE = 27) have higher error rates.

considered healthy, while those with scores below 27 are classified as dementia-related. Therefore, it is normal to observe a higher misclassification rate around this threshold, as the model may find it challenging to distinguish between healthy and dementia-related samples that are close to the boundary.

3.5 Conclusion

In this chapter, we have explored the use of autoencoders and large language model representation for anomaly detection in dementia-related data. We conducted experiments with different configurations, including different embedding types and various autoencoder variants such as deep autoencoders, denoising, and variational autoencoders. The results demonstrated that the choice of architecture and latent dimension significantly impacts the performance of the autoencoder in detecting dementia-related anomalies. Our findings indicate text embeddings from large language models, such as RoBERTa and LLaMA, can effectively capture the nuances of dementia-related data. These embeddings, when processed through autoencoders, can help identify anomalies in the data, providing a valuable tool for early detection and diagnosis of dementia.

Chapter 4

Detecting Dementia from Face-related Features with Automated Computational Methods

4.1 Introduction

Dementia can significantly impact a person's ability to express themselves, both verbally and non-verbally. As the disease progresses, individuals may exhibit changes in their facial expressions, such as reduced emotional responsiveness, flat affect, or difficulty in recognizing and interpreting social cues. These changes can make it challenging for caregivers and healthcare professionals to assess the patient's condition accurately [145]. In addition to cognitive and behavioral changes, dementia can also affect a person's facial features. Research has shown that individuals with dementia may exhibit distinct facial characteristics, such as reduced facial muscle tone, changes in skin texture, and alterations in facial symmetry [146]. These changes can be subtle and may not be immediately noticeable, but they can provide valuable insights into the progression of the disease. Apathy is a common symptom of dementia, characterized by a lack of interest, motivation, and emotional responsiveness [147]. Individuals with dementia may appear indifferent or emotionally flat, which can make it difficult for caregivers to engage with them effectively. This apathy can also manifest in reduced facial expressions, making it challenging to assess the patient's emotional state and overall well-being [145].

Facial expression analysis has traditionally been a challenging task due to the complexity of human emotions and the variability in individual facial features. For example, the action units (AUs) defined in the Facial Action Coding System (FACS) [148] provide a comprehensive framework for categorizing facial movements, but they require manual annotation and interpretation, which can be time-consuming and subjective. Recent years have seen significant advancements in computer vision and machine learning techniques, enabling more automated and accurate analysis of facial expressions [149]. These methods can assist detecting faces and extracting relevant features, such as facial landmarks and AUs, which can then be used to classify emotions or detect changes in facial expressions over time. The development of deep learning models, particularly time-series models that can analyze sequential data, has further strengthened the potential of detecting dementia from videos of dementia patients.

Compared to the extensive research on speech and language-based dementia detection, facial features have received much less attention, even though facial features have proven to be effective indicators in dementia detection [64]. Previous works [63, 64] have made some explorations on dementia detection using facial features or facial images. However, these works only perform experiments with raw facial images or some facial features. The dataset used in their experiments is relatively small and limited, which could also introduce bias.

4.2 Related Work

It is known that demented people usually have problems with memory and language because dementia damages the part of the brain that controls the functions of languages. Many researchers paid attention to the speech data and many methods were developed to distinguish dementia patients from healthy subjects [124, 150–154]. These methods mainly used NLP techniques to evaluate the data. Different features were extracted and different types of machine learning models were proposed to perform classification on the text data, with labels indicating whether the text samples come from a dementia patient or a healthy subject. Weiner et al. [150] compared two feature extraction pipelines for dementia detection. The first pipeline involves manual transcriptions, while the second pipeline uses transcriptions created by automatic speech recognition (ASR). According to the study, the transcription quality of the ASR system is a dependable feature for detecting dementia on its own. Moreover, the features extracted from the automatic transcriptions exhibit comparable or slightly superior performance when compared to those derived from manual transcriptions. Mirheidari et al. [151] investigated how word vector representation models perform in detecting signs of dementia. Motivated by the fact that dementia patients suffer from impairment on accurately expressing something, they analyzed conversations designed to test examinees' long-term and short-term memory. Three methods were proposed to show the potential of word vectors in a classification task. Their study concluded that it is possible to detect signs of dementia with a speech recognizer even though the recognition result has a high error rate. Zhu et al. [152] found

that transfer learning models utilizing text data outperform those using audio data, likely due to the high similarity between the pre-training text dataset and the Cookie-theft picture text dataset. Multi-modal transfer learning shows a slight improvement in accuracy, indicating that audio and text data provide limited complementary information. However, multi-task transfer learning results in limited improvements in classification and a negative impact on regression. They also identified that inconsistencies between AD/non-AD labels and Mini-Mental State Examination (MMSE) scores can limit the performance of multi-task learning. Mahajan et al. [153] re-implemented NLP methods that utilized CNN and LSTM architectures and specific conversational features to diagnose AD. They examined why these models had lower accuracy on the ADReSS [155] dataset than the DementiaBank [115] dataset. Additionally, they created a deep learning-based solution using Recurrent Neural Networks (RNNs) that performed binary classification of Alzheimer's Dementia from natural speech. Their method combined acoustic features using the Speech-GRU, improving accuracy by 2% over the acoustic baseline. Furthermore, when enriched with some targeted features, their method outperformed acoustic baselines by 6.25%. They proposed a bi-modal approach to AD classification and discussed the potential benefits of their method. Shibata et al. [154] adapted and applied the idea density to the Japanese language. They defined several rules to count the idea density in Japanese speech and proposed a method to evaluate the idea density using machine translation. They proposed different pipelines for estimating the idea density, and compared the performance of different idea density calculation methods. They concluded that these findings prove that it is feasible to apply the method in the area of dementia detection. Farzana et al. [89] evaluated how removing the disfluency in the annotated result of an automatic speech recognition system affects the performance of dementia detection. They used an existing tool that detects and tags disfluency in transcriptions of speech. Through a set of experiments, they found that removing the disfluency has a bad influence on dementia detection performance, which reduces the detection accuracy.

Ilias et al. [156] proposed a novel method that combined two modalities such as speech and transcripts into one model based on the vision transformer. They used the gated multimodal unit to control how much influence each modality will have on the final classification. They also used crossmodal attention that learns the relationships between modalities in an effective way. They tested their work on the ADReSS Challenge dataset and proved their model's superiority over existing methods.

The above-mentioned most recent researches concentrate on the use of speech and text data to detect dementia. In some earlier works, researchers also explored the potential of facial expressions in dementia detection. Asplund et al. [157, 158] observed noticeable differences in facial expressions among individuals with dementia and demonstrated the potential to use these expression patterns for identifying the condition. Their earlier study [157] revealed that some patients with dementia tend to exhibit fewer facial expressions. By applying the Facial Action Coding System (FACS) under conditions of both positive and negative emotional stimuli, they identified a diminished ability in these patients to generate facial expressions. In a subsequent investigation [158], the authors compared two distinct methods for interpreting facial cues in dementia patients, revealing discrepancies between them. Overall, their findings indicated that individuals with dementia often present facial expressions with reduced frequency or clarity, suggesting that such facial indicators could potentially function as diagnostic markers.

In more recent efforts, researchers have expanded on these early findings by collecting and analyzing facial expression data from dementia patients [63, 64, 159]. These works seek to address the relative lack of attention paid to facial cues in dementia research compared to more extensively studied speech and language features. One study, for example, introduced an automated approach to identifying dementia by analyzing facial features captured during spoken conversations with 24 participants. Utilizing L1-regularized logistic regression to classify subjects, the model achieved an accuracy of 0.82. Key discriminative features included lip movements, AUs, and eye gaze behavior. These results underscore the viability of using facial behavior data captured during dialogue for dementia screening, offering practical support for healthcare professionals.

Umeda et al. [63] further investigated the capacity of artificial intelligence to differentiate between individuals with and without cognitive impairments based solely on facial images. Drawing on data from 121 cognitively impaired and 117 cognitively intact individuals, the researchers tested five deep learning architectures. The Xception model, optimized using Adam, delivered the highest performance, achieving 87.31% sensitivity, 94.57% specificity, and an overall accuracy of 92.5%. A significant association was found between the computed “Face AI score” and cognitive decline, suggesting that facial features could be developed into biomarkers for dementia diagnosis. Despite these promising results, the authors also acknowledged a major limitation: the risk of institutional bias. As noted in their study, the model might inadvertently learn environmental features specific to the institution rather than markers truly indicative of dementia.

Several studies have examined the facial expressions of dementia patients in controlled settings, although these analyses were not applied to dementia detection directly: Liu et al. [159] investigated the effects of various sound clips on the emotional responses of elderly dementia patients in a care facility using facial expression recognition software. The study observed temporal changes in emotional reactions as expressed through facial expressions. Compared to a silent condition, participants exhibited significantly higher facial valence in response to the three sound-based interventions. Each sound treatment produced distinct effects on indices related to fear, disgust, and melancholy. Additionally, the timing of sound exposure relative to the start of activities was found to influence both valence and arousal levels. Gender and age were also associated with variations in the valence of facial expressions.

Jiang et al. [160] employed a computer vision-based deep learning algorithm to objectively analyze facial emotions in a large cohort of participants ($n = 493$) with varying levels of cognitive impairment, including healthy control subjects, during a memory task. The results revealed distinct facial expression patterns associated with different stages and etiologies of cognitive decline. Compared to cognitively healthy individuals, those with impairment displayed more negative emotions, fewer expressions of happiness, and greater overall facial expressiveness. Although the approach did not achieve reliable

differentiation between underlying etiologies, the findings suggest that facial expression analysis could serve as a useful passive screening tool for cognitive impairment.

As the above introduced works show, many researches on dementia detection paid attention to the exploit of speech and text data, especially the DementiaBank dataset and its derivative the ADReSS dataset [155], which have been extensively studied and investigated. However, some researchers found that patients who have dementia behave differently from healthy subjects in facial expressions [157, 158]. At the same time, the video data is a rich and informative source where we can easily extract facial features. Motivated by these facts, we decided to explore the potential of facial features in dementia detection.

The use of facial features to detect dementia offers several advantages over other screening methods. The traditional diagnosis of dementia often involves a battery of tests, including brain scans and psychiatric evaluations, which can be stressful and time-consuming for patients. Many patients may be reluctant to undergo these tests, which can lead to delays in diagnosis and treatment. In addition to psychiatric evaluations and imaging scans, speech-based detection of dementia has emerged as a prominent area of investigation. However, gathering speech data often necessitates the use of the Boston Cookie Theft Test, a well-established tool for assessing cognitive function in individuals with suspected dementia. To collect speech data, patients are required to actively and intentionally describe the scene depicted in the picture, a process that typically requires the involvement of trained instructors. As such, this approach can still be costly and time-consuming, impeding the large-scale collection of speech data for diagnostic purposes. The diagnostic process for dementia can be particularly challenging, with patients often expressing anxiety and confusion about the tests they are undergoing.

While traditional methods such as psychiatric tests or brain scans can be stressful for the patients, the process of collecting facial features can be smoother and easier for the patients. It does not require the patients to undergo difficult tests which may be stressful for dementia patients who might feel reluctant and incompetent to fulfill such tests. Furthermore, facial feature data can be collected concurrently with other tests or screenings,

eliminating the need for additional appointments and reducing the burden on patients. Facial analysis can not only function as a stand-alone biomarker for dementia detection but can also be used in conjunction with other types of data, such as speech or brain scans, to provide a more comprehensive assessment of cognitive function. This complementary approach may lead to more accurate and reliable diagnoses, ultimately improving patient outcomes. Therefore, in this chapter, we are motivated to investigate the potential of facial features in dementia detection.

In this Chapter, we first examined and investigated the institutional bias problem, which exists in but is not limited to the PROMPT (PROMPT for Objective Measures using computational Psychiatry Technology) dataset ¹ [161]. We employed Grad-CAM [162] to visualize the bias in the dataset and conducted a series of cloze tests to validate the existence of the bias. After we confirmed the existence of the institutional bias, we proposed to use facial features such as face mesh, Histogram of Oriented Gradients (HOG) and AUs to detect dementia. Compared with raw face images, these features, especially face mesh and AUs, can remove the bias introduced by different lighting conditions or different environments in data collection. They can also anonymize the data to protect patient privacy. The PROMPT dataset consists of more than 60 hours of video data. With such a relatively large dataset, we conducted experiments to investigate the effectiveness of the facial features mentioned above.

The main contributions of this chapter are as follows:

- We examined and confirmed the existence of institutional bias in the PROMPT dataset. We proposed to use Grad-CAM to visualize the bias in the dataset and conducted a series of cloze tests to validate the existence of the bias. We also proposed several methods to remove the bias in the dataset.
- We investigated and analyzed the potential of using face mesh, HOG features and AUs in dementia detection. These features have many advantages, such as the ability to remove the bias in the dataset and anonymize the dataset to protect patients' privacy.

¹<https://www.i2lab.info/resources>

- We proposed several methods and pipelines to normalize and process these features and to detect signs of dementia using the facial features.
- We adapted and improved the method in [64] through segmentation and voting. We validated this method on a relatively large dataset – i.e., PROMPT dataset.

The structure of the chapter is organized as follows this: Section 4.3 introduces the dataset and the features employed in the experiments. Section 4.4 explains the settings and parameters of the experiments. Section 4.5 shows the results of the experiments of different features under different methods. Section 4.6 gives the conclusion and future directions.

4.3 Dataset and Facial Features

4.3.1 PROMPT Dataset



Figure 4.1. Camera and mic settings used in the data collection process.

The Keio Medical School collected voice and video data from participants with various diagnoses, including healthy persons, patients with dementia, patients with bipolar disorder, and patients with depression, for the PROMPT dataset [161]. Only videos of dementia patients and healthy subjects will be used for this specific study because our focus is on detecting dementia by analyzing facial expressions.

Table 4.1. Demographic information of subjects by age and MMSE Score

Age	Count	MMSE	Count
20–29	3	0–5	2
30–39	3	6–10	4
40–49	4	11–15	5
50–59	1	16–20	16
60–69	23	21–25	14
70–79	26	26–30	56
80–89	33	Unknown	20
90–99	2		
Unknown	22		

In the collection process, the dementia patients participated in several interview sessions conducted by professionals, with at least a 2-week gap between sessions. During each session, the patients were required to undergo several interviews. The video and audio recordings of the sessions were captured separately from two microphones and an Intel Realsense camera. One microphone was set at the patients' place and another set at the doctor's place. The camera was set on the table in front of the patients, with a distance of about 1 meter between them as shown in Fig. 4.1.

- **Free Speech:** Within the free talk task, individuals engaged in casual discussions with the interviewer. Despite its being named “free talk”, the interviewers guided the conversation by posing casual daily questions from a predetermined list. These included questions like “Did you have a nice sleep last night?” and “Did you feel too hot or too cold during sleep?”. This part of the data emphasizes intense interactions that are short and quick between the subjects and the psychologists. The subjects were replying to the psychologists immediately and casually.
- **Question and Answer (Q&A):** In the Q&A task, participants responded to a series of three sequential queries posed by the interviewer as follows: 1) What would you do if there was a fire? 2) What would you do if you lost the umbrella you borrowed? 3) What would you do if you found an envelope with an address on it? These

questions establish relatively complex scenarios in daily life. By interrogating the subjects with these questions, we expect to see different responses from dementia patients and healthy subjects.

- **Picture Description:** In this part of the task, the participants were instructed to describe the content of an image displayed, which showed a scenario with a boy and two girls having an argument over food.

The video data collected in the free talk task was used as the dataset in the experiments. Patients are recorded using a Real Sense camera placed on a table in front of the patients, with the upper body of the patients captured in the videos. The videos are stored as a sequence of images in “*.JPEG” format. In total, we have 250 healthy samples collected from 61 subjects, and 195 dementia samples collected from 56 subjects. The demographic information of the subjects are shown in Table 4.1. Despite that each subject may have multiple video recordings, we split the dataset into training and testing sets by subject, i.e., all video recordings of a subject are either in the training set or in the testing set, but never both. This is to ensure that the model does not learn from the same subject in both training and testing sets, which could lead to overfitting and biased results. A single subject may undergo the interview several times and provide more than 1 video recording. Each video recording lasts approximately 10 minutes. We extracted the face mesh data, HOG features, and AUs from the video data. These features are described and explained in the following sections.

4.3.2 Raw Face Images

Raw face images are the original images captured by the camera during the data collection process. It is intuitive to use raw face images for dementia detection, as they contain rich information about the facial expressions and features of the subjects. However, as data scarcity and lack of diversity are common problems in dementia detection area, using raw images is risky because the model may learn the bias introduced by the data collection process, such as lighting conditions, camera angles, and backgrounds. Several studies

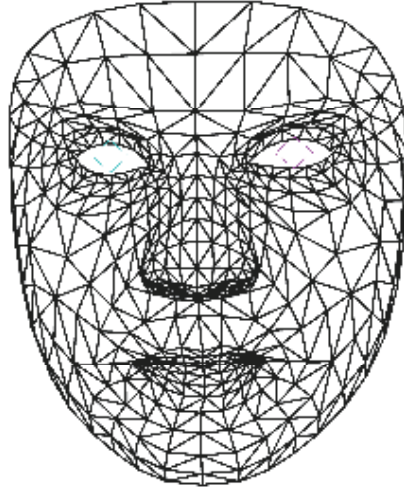


Figure 4.2. An example of face mesh extracted using MediaPipe.

have explicitly pointed out the existence of institutional bias in their datasets, for example, Liu et al. [62] reported institutional bias in the commonly used Dementia Bank audio dataset, reporting nearly perfect performance using silent segments of the audio. Umeda et al. [63] also pointed out the existence of institutional bias in their dataset.

These studies encourage us to investigate the existence of institutional bias in the PROMPT dataset and to propose methods to remove the bias in the dataset. Therefore, despite the fact that raw face images are intuitive and easy to use, we do not use them directly in our experiments, instead they will be used to conduct a series of cloze tests to validate the existence of institutional bias in the dataset. In the conducted cloze tests, we systematically and progressively removed the background and other irrelevant information from the raw face images, and observed how the classification performance changes with different levels of content removal.

4.3.3 Facial Feature: the Face Mesh

A face mesh is a digital representation of the geometry of a human face, typically created using 3D computer graphics techniques. It consists of a collection of interconnected vertices, or points in space, that are arranged in a mesh-like pattern to approximate the shape of the face. The face mesh is often used in computer vision and graphics applications, such as facial recognition, animation, and virtual reality, to track and manipulate the movements of the face in real time. By using a face mesh, it is possible to capture and analyze the subtle changes in facial expressions that convey emotion and communication cues, which can be useful in various fields such as psychology, marketing, and entertainment. With face mesh, it is possible to create a detailed 3D model of a patient's face without capturing any personally identifiable information, such as facial features or skin color. This allows medical professionals to analyze and track facial movements and expressions without compromising patient privacy. Additionally, the use of face mesh can be particularly useful in telemedicine applications, where patients may not be able to visit a healthcare provider in person. By using face mesh technology, doctors and other healthcare professionals can remotely assess a patient's condition and make informed diagnoses, all while ensuring patient privacy and security. Overall, the use of face mesh in medical applications offers a powerful combination of accuracy and privacy protection, making it an ideal tool for healthcare professionals who want to deliver the best possible care while preserving patient confidentiality.

MediaPipe [163] uses a combination of machine learning and computer vision techniques to extract a detailed face mesh from an input video stream or sequences of videos. The face mesh extracted by MediaPipe consists of a total of 478 vertices that are located at key points of a human face from each frame. These vertices are interconnected to approximate the 3D facial surface of the face in the given image. An example of a face mesh extracted by MediaPipe is shown in Figure 4.2.

4.3.4 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) is a widely utilized feature descriptor in computer vision for detecting objects. It extracts information from an image by dividing it into small cells and computing a histogram of oriented gradients within each cell. This process captures data about the local edge orientation and magnitude, which are then used to describe the appearance of an object. The HOG descriptor is able to capture these features without being affected by changes in illumination and provides a compact representation of the object that can be utilized for object classification or detection. The HOG feature descriptor has been applied with great success in various fields, including but not limited to pedestrian detection, face recognition, and object tracking.

HOG has been shown to be a useful tool in medical imaging analysis, specifically in identifying and detecting various structures of interest, such as tumors or lesions [164, 165]. By extracting information from images, HOG can provide a unique feature descriptor that can help distinguish between different types of structures. Additionally, HOG is invariant to changes in geometric and photometric transformations, making the HOG a robust feature.

4.3.5 Facial Feature: Action Unit

AUs are a standardized set of facial muscle movements used to describe and quantify different facial expressions [148]. Developed by psychologists Paul Ekman and Wallace Friesen in the 1970s, AUs offers a systematic and objective approach to measuring changes in facial expressions. There are currently 43 AUs that correspond to specific muscle movements, and they are widely applied in fields such as psychology, neuroscience, and computer vision to study emotional expression, communication, and social interaction. The use of AUs in analyzing facial expressions can provide valuable insights into how people perceive and respond to emotions [166]. Moreover, AUs have been incorporated into technology such as facial recognition software and virtual assistants, enabling these systems to interpret human emotions and provide a more natural and empathetic user experience.

Each AU is represented by a number, which corresponds to the intensity of the movement on a scale from 0 to 5, where 0 means the muscle is at rest and 5 means the muscle is fully contracted. The values of each AU can be combined to describe different facial expressions, such as a smile, frown, or surprise. The AUs are formulated based on the anatomical location of the muscles involved in the movement. For example, AU 12 corresponds to the contraction of the zygomatic major muscle, which is located in the cheek and lifts the corners of the mouth, resulting in a smile. Another example is AU 2, which corresponds to the contraction of the frontalis muscle, located in the forehead, and is associated with raising the eyebrows in surprise or confusion.

4.4 Experiment Designs and Settings

In the following sections, we will explain the design and settings of our implementation. We used different processing techniques, models, and cross-validation schemes for different types of features to optimize the classification performance. In the following parts, the experiment settings for each type of feature are described and explained, respectively.

4.4.1 Cloze Test Experiments on Institutional Bias

We conducted a series of cloze tests to validate the existence of institutional bias in the PROMPT dataset, by progressively masking the background and other irrelevant information in the raw face images using multiple tools such as OpenCV [167] and PaddleHub [168], and MediaPipe [163]. We implemented a CNN model to perform the classification and used Grad-CAM [162] to visualize the bias in the dataset. In total, we conducted four experiments with different pre-processing pipelines to remove the background and other irrelevant information in the raw face images.

In Experiment I, we did not pre-process the images and simply used the raw data set to train a simple CNN and perform the classification. The network contains only 5 convolutional layers with Rectified Linear Unit (ReLU) Activation and a pooling layer of

size 2. The kernel size of each layer is 3 and the input/output channel number of each convolutional layer except the first one is 64.

In Experiment II, we use PaddleHub [168] to mask the background, which was shown to have a dominant influence on the classification, and use OpenCV [167] to crop the face part of the image. This created a new sub-set instance where the background of the images is replaced by white pixels. We applied the same classifier as in Experiment I to this sub-set instance.

In Experiment III, we first use MediaPipe [163] to extract the face mesh of each frame. Following the contours provided in the face mesh, we mask the image except for the face area. After that, we use OpenCV to crop out only the unmasked area of the image. This creates a sub-set instance that contains only the area of the face, without any background information. We performed classification again using the same network as in Experiment I on this sub-set instance.

In Experiment IV, we implemented experiments to evaluate how different cropped areas (eyes, nose, mouth) of the face could influence the results of the classification. We implemented the experiment using the same network as described in Experiment I on these data set instances and evaluate how different areas of the face have different importance in dementia detection. The Prompt data set collected is in the RGB format with full color. We tried converting the RGB images into grayscale images and see how color information could affect the the outcome of the classifier in this experiment. Fig. 4.3 shows the overview of the four cloze test experiments on the raw face images.

4.4.2 Experiment Settings for 2D Facial Features

To validate the effectiveness of 2D facial features, we implemented a baseline experiment by applying the method proposed by Tanaka et al. [64] on our datasets using the same data split to make a fair comparison. In [64], the authors proposed a method to detect dementia by analyzing facial landmarks, action units, and response times extracted from the video data. Following their method, we extracted the following facial features using OpenFace [169]:

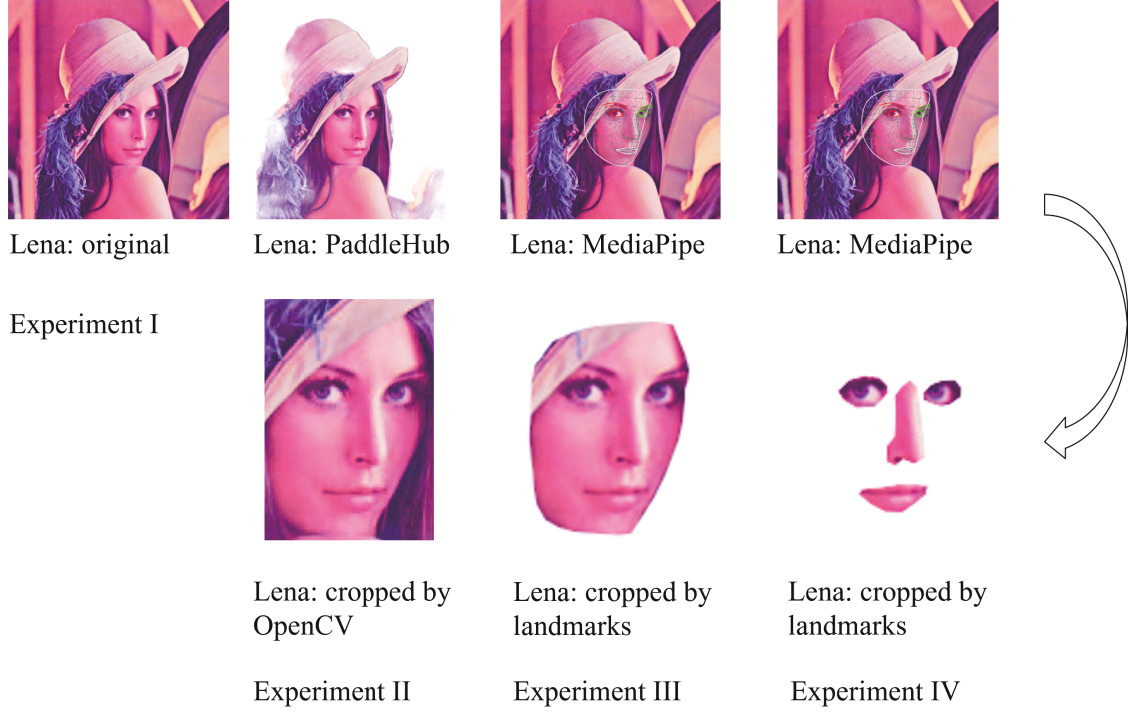


Figure 4.3. The cloze test experiments on the raw face images. The first row shows the original images, and the second row shows the masked images.

- 2D facial landmarks (mouth region: points 48–67),
- Face pose (x, y, z),
- Gaze angles (x and y),
- Action unit (AU) intensities: AUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, 45,

However, we did not include the response time in the feature set, as it is not available in the PROMPT dataset. In the dataset used in [64], they prepared 3 pre-defined questions for the subjects to answer. In their data collection, these questions were uttered by a computer agent, therefore, the response time can be calculated by measuring the time between the computer agent’s question and the subject’s answer. However, in the PROMPT dataset, the data collection process is more flexible, and the subjects are having a relatively free conversation with the psychologists. Therefore, the response time cannot be calculated in the same way as in [64]. We have replicated their detecting method on our

dataset. Besides, we have also analyzed the importance of different facial features in the classification task.

4.4.3 Experiment Settings for Face Mesh

We utilized MediaPipe to analyze the face mesh data from the 447 videos in the PROMPT dataset’s free speech section. These videos typically have a duration of around 10 minutes and a frame rate of 30 frames per second. MediaPipe labels each frame in the video with information about the face mesh. Consequently, we converted each video into a collection of face mesh data consisting of 18,000 frames. Each frame of face mesh data includes 478 landmarks, which are annotated with (x, y, z, v) values. Here, (x, y, z) represents the three-dimensional coordinates of each landmark, and v indicates whether the landmark is visible or not.

MediaPipe does not adjust or standardize the positioning or size variations of the face mesh data obtained due to differences in camera settings, such as angles or distances. For each input image, the face mesh was annotated at the exact location where the face locates in the image. Therefore, to reduce the influence of such factors on the classification, we normalized the face method in the following way:

First, coordinate transformation is applied to the coordinates of every landmark of a face mesh. In the extraction, the face in the original image and the annotated face mesh share the same coordinates. The faces can possibly appear anywhere in the image, and so does the face mesh. To make the LSTM network better learn the patterns of the face mesh, we remap the coordinates of each landmark by transforming them with respect to a specified landmark in the center of the face. It makes this landmark the new origin of the coordinator and the rest of the coordinates are depicted with the new origin. After the coordinate transformation normalized the movement of the face mesh, we resize the coordinate range of the face mesh to $[-1, 1]$. As the videos sometimes are recorded at different distances from the patient, the size of the face in the video may vary due to perspective. Resizing the coordinate range ensures that the face mesh of each

individual data is always kept at the same size. An example of face mesh before and after normalization is shown in Figure 4.4.

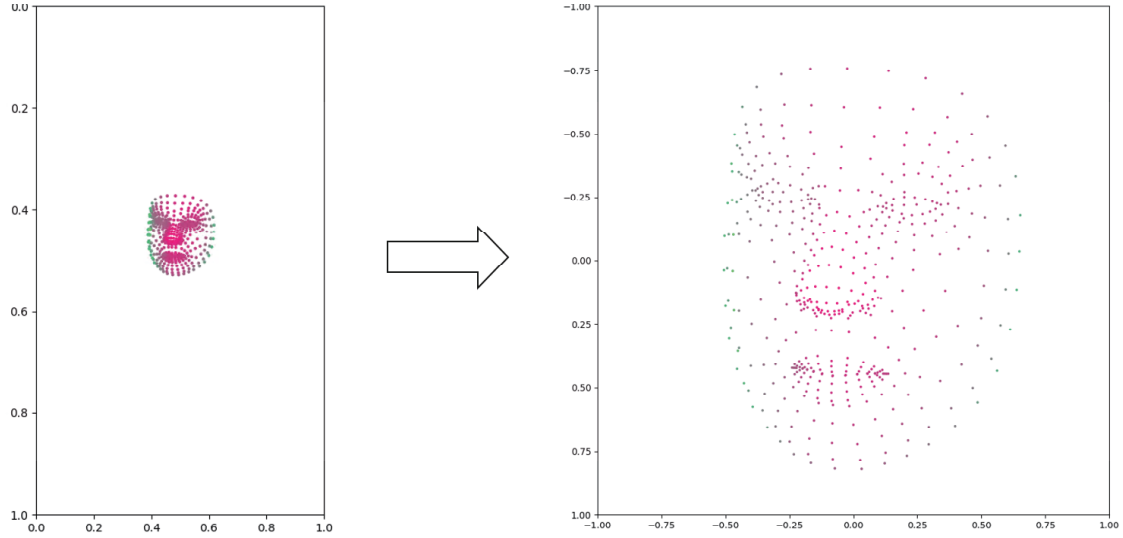


Figure 4.4. An example of face mesh before and after normalization.

Even after removing the visibility from the data, each frame still has 1434 features, which is a quite large feature size. However, these 1434 features share a lot of correlated information and internal connection because face muscles do not move independently but usually move as a group. Besides that, each muscle is annotated to multiple landmarks, which further strengthens the correlation among the landmarks. Therefore, instead of using the normalized face mesh data to train the machine learning model, we use PCA to reduce the size of the data and perform classification.

PCA is a widely used statistical technique that helps in understanding and simplifying complex data sets. It provides a way to reduce the dimensionality of data while preserving as much information as possible. By extracting the most significant patterns and relationships within the data, PCA enables us to uncover the underlying structure and identify the key variables that drive the variation in the dataset. Through this process, PCA aids in visualizing and comprehending large datasets, enhancing data exploration, feature selection, and machine learning tasks. The core idea behind PCA is to find a new set of variables, called principal components, which are linear combinations of the original variables. These principal components are derived in such a way that the first

component captures the maximum amount of variation present in the data, the second component captures the remaining variation orthogonal to the first component, and so on. Each subsequent component accounts for the maximum remaining variation while being uncorrelated with the previous components. In this experiment, we applied PCA to process the extracted face mesh data before training the classification model.

An LSTM network with a hidden size of 32 was applied as the classifier to perform dementia detection based on the face mesh data. Each frame of the face mesh data was first reduced by PCA into 16 principal components, which are forwarded and input into each time step of the LSTM network in time order. After the LSTM network has processed the last frame of the face mesh data, the hidden states at the last time step of the LSTM network are forwarded to a dense layer. The dense layer takes the hidden states of the LSTM network as its input and outputs a score pair as the probability of the sample belonging to either dementia or non-dementia. We used the cross-entropy loss as the loss function and the Adam optimizer with a learning rate of 0.001 to train the model.

To validate the performance of our proposed methods, we created a test set from the PROMPT dataset and made sure that:

- The identity information of each patient was not leaked: in the data collection of the dataset, each patient paid different times of visits to the doctors. Therefore, each patient provides 1-10 samples in the dataset. In the creation of the test dataset, if a patient is in the test set, we will put all his or her data in the test set.
- The number of samples and the number of subjects of each category are approximately equal: because of the fact mentioned above, we also managed to keep a balance in each category of the test set.

After completing the data pre-processing and PCA reduction, we trained the LSTM network on the training data, and the performance was evaluated on the test set.

To address and mitigate the inconsistency in the postures and angles of the patients in the videos, besides for normalizing the face mesh data, it is also possible to add regularization to the model by adding more variations to the training data. Therefore, by inverting

the process of the normalization, we can generate more training data by add random shift, random rotation, and random scaling to the face mesh data. This process is similar to data augmentation, which is widely used in computer vision tasks to improve the robustness of the model. In this experiment, we applied random shift, random rotation, and random scaling to the face mesh data to generate more training data.

Our proposed method is novel to address the problem of low-quality data in video datasets by converting the data into face mesh and manually editing them in 3D space. The video data in dementia detection tend to have lower quality due to several reasons:

- Physicians or psychiatrists are not familiar with machine learning, and factors that cause degradation may be ignored by them, such as the setting of camera, the device used, clothes of the patients, etc., which may not be strictly controlled.
- Patients suffer from cognitive impairment and cannot follow the instructions, leading to inconsistent postures and angles in the videos.
- Senior citizens face barriers in commutation and cannot attend data collection in a specified place, leading to variations in the video data.

Our proposed method provides a powerful tool to solve the above challenges in low-quality video datasets. The experiments conducted in this section demonstrate the effectiveness of our method in detecting dementia from the PROMPT dataset, which contains videos with low quality and variations in postures and angles.

Table 4.2 summarizes the comparison of the different facial features, including MediaPipe Face Mesh, HOG, Haar Cascades, and Facial Landmarks. MediaPipe Face Mesh provides a high-dimensional representation of the face with 468 3D landmarks, capturing detailed facial expressions and head movements. The 3D estimation and track allows us to edit, transform and manipulate the face mesh data, to overcome the inconsistencies in the postures and orientations of the patients in the videos. Besides, the rich information content of the face mesh data also has the potential to distinguish dementia patients from other diseases, such as depression or Parkinson’s disease, which cause similar problems in facial expressions. It is robust to institutional bias, as it avoids the influence of lighting,

Table 4.2. Comparison of the different facial features: MediaPipe Face Mesh, HOG, Haar Cascades, and Facial Landmarks.

Feature Type	Dimensionality & Resolution	Information	Robustness to Institutional Bias
MediaPipe Face Mesh	High (468 3D landmarks)	Very high, captures expressions and head movements	Avoid light, background, and posture influence
HOG	Low (2D feature)	Low, sensitive to noise; lacks semantic localization	Avoid background
Haar Cascades	Very low (2D feature)	Very low, binary features without geometry	Sensitive to illumination
Facial Landmarks (68-point)	Medium (68 2D landmarks)	Moderate, can track gross expressions	Sensitive to occlusion caused by posture

background, and posture variations. In contrast, HOG features are low-dimensional 2D features that capture local gradients but are sensitive to noise and lack semantic localization. HOG features can be affected by background variations, making them less robust to institutional bias. Haar Cascades, on the other hand, provide very low-dimensional binary features without geometric information, making them very low in information content and sensitive to illumination changes.

4.4.4 Experiment Settings for HOG features

We utilized the Open Face framework [169] to extract HOG features from video data in the PROMPT dataset. The Open Face approach involves several sequential steps, as outlined in [170]. First, it performs face tracking using the Constrained Local Neural Field (CLNF) method [171] as a facial landmark detector and tracker to track the face and obtain geometric features. Then, the alignment and masking step is responsible for rectifying the tracked face and compensating for variations caused by plane rotation and scaling. After the face is aligned, appearance features are extracted in the form of HOG features.

This involves dividing the face into blocks of 8×8 pixels, resulting in 12×12 blocks with histograms of 31 dimensions each. Collectively, these histograms form a 4464-dimensional vector that describes the face’s appearance, which will be used as features in this section. Additionally, geometry features are extracted, resulting in 227-dimensional vectors. Support Vector Machines (SVM) and Support Vector Regression (SVR) are then employed to predict the occurrences and intensities of AUs based on the HOG features and geometry features. In this section, we will explain the use of HOG features, and leave AUs for the next section.

We used the Open Face tool to extract HOG features from each frame of the videos in the PROMPT dataset. We adapted the model in Section 4.4.3 by tuning the parameters of the model to fit the HOG feature format. We used the same test described in Section 4.4.3. Each frame of the HOG features was input into an LSTM network in time steps, and the final hidden states after the LSTM model has processed the whole HOG feature sequence are passed to a dense layer for final classification. Since HOG features are based on local gradients, changes in lighting conditions and color can alter the gradients and subsequently impact the extracted features. Different lighting conditions can lead to variations in the appearance of edges and shapes, potentially resulting in different gradient orientations and magnitudes. This, in turn, can influence the HOG feature representations, making the HOG features likely to be biased by the lighting conditions. Due to this reason, we want to compare how a single frame of HOG features can detect dementia and how multiple frames of HOG features can detect dementia. This may reveal how much the HOG features were affected by the bias in the images. Therefore, we trained several networks while varying the number of frames used in the training.

4.4.5 Experiment Settings for Action Units

Open Face utilizes machine learning models to extract 17 intensities of AU from each frame of the videos in the PROMPT dataset. The process for extracting these AU intensities using machine learning models is described in Section 4.4.4. To analyze the temporal behavior of these AU intensities, we follow the approach presented in [64]. We calculate

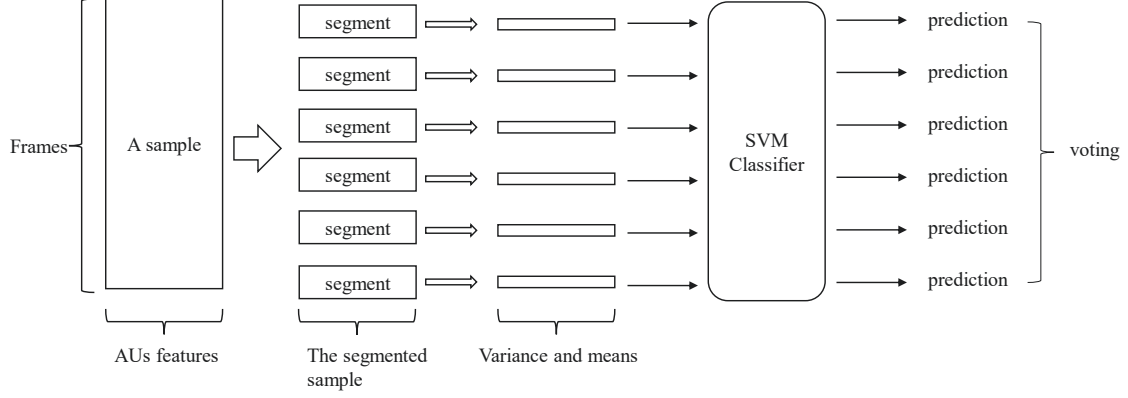


Figure 4.5. The overall pipeline of performing dementia detection using AUs.

the mean and variance of the 17 AU intensities over time for the entire video, which typically spans approximately 10 minutes. This calculation yields a feature vector consisting of 34 elements for each video in the PROMPT dataset. However, the PROMPT dataset consists of 447 videos from 117 subjects. Each video lasts around 10 minutes, with a frame rate of 30 frames per second (FPS). Consequently, we have approximately 18,000 frames per video, and each frame contains 17 AU intensities. Merely reducing the dimensionality from over 300,000 features to 34 features would discard a significant portion of the video’s information. To address this issue, we employ a segmentation method that divides each video into smaller video clips with a fixed number of frames. This segmentation allows us to generate additional features when computing the mean or variance from the AU intensities. In this study, we explore two segmentation settings, where we divide the videos into either 1024 frames or 512 frames to capture different temporal contexts and increase the number of features available for analysis.

After extracting the mean and variance features from the PROMPT dataset, we used an SVM classifier with RBF kernel to classify the video clips into either dementia or healthy category. Since training an SVM with AU data is much faster than training an LSTM, we ran a Leave-One-Subject-Out (LOSO) scheme in the experiment. Each time, we left out all video clips of a subject as the test set, used the remaining videos as training data, and trained an SVM classifier to perform classification. After training the model, it was used to make predictions for each held-out video clip. Among all the video clips of the held-out subject, we implemented a voting system to make the final decision. Also, in

this experiment, we used the mean and variance of the AUs to perform the classification. The pipeline of performing dementia detection using AUs is shown in Figure 4.5. Figure 4.6 shows the difference between healthy subjects and dementia patients in terms of the mean and variance of the AU index [2, 4, 5, 7, 8, 9, 10, 12, 14, 15, 16]. We noticed from this figure that not all AUs have a significant difference that includes the indicators for dementia. Therefore, we also tried using only 11 AUs that show a significant difference in the experiment to detect dementia and discarding AUs that could mislead the classifier.

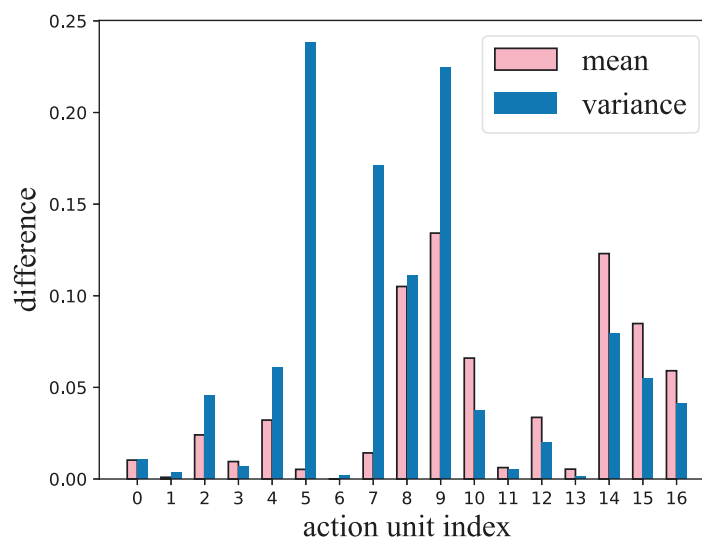


Figure 4.6. The difference of variance and mean of each action unit between healthy subjects and dementia patients.

Table 4.3. Cumulative explained variance for the top 1 – 16 principal components.

Component number	cumulative explained variance
1 component	0.33
2 components	0.55
3 components	0.72
4 components	0.81
5 components	0.90
6 components	0.94
7 components	0.98
8-16 components	0.99

4.5 Results and Evaluation

4.5.1 Cloze Test Experiments on Institutional Bias

Exp. I Classification with unprocessed image data

In Experiment I, we used a simple CNN to classify dementia patients and healthy subjects, achieving 89.46% accuracy using raw images as the input. Suspecting potential data bias, we used CAM to examine the model's decision process. Fig. 4.7 shows a heat map generated by CAM (Despite we masked the patients' face for protecting privacy in the dissertation, they were not masked in the training or testing process). In the heat map, red areas contribute positively to the specified class, while blue areas do not.

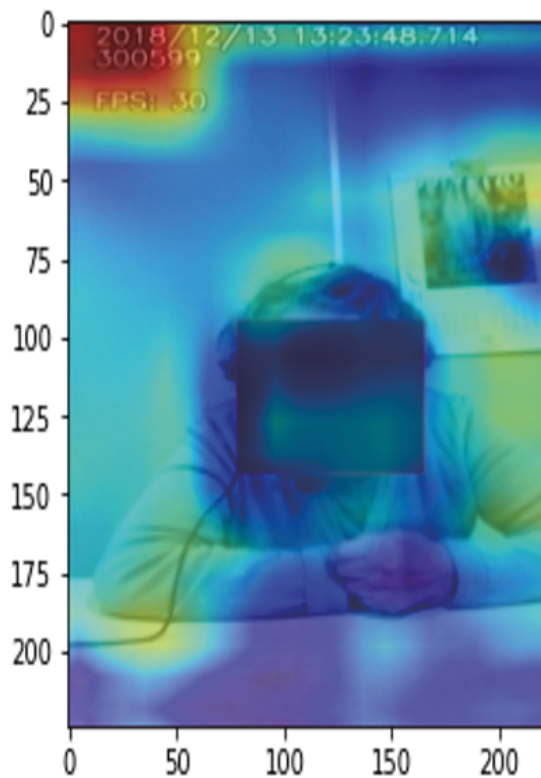


Figure 4.7. An example of CAM heat map that shows the focusing area of the CNN model.

For binary classification, CAM generates two heat maps per class. In Fig. 4.7, the specified class is dementia, with the red circle indicating why the model classifies the image as from a dementia patient. We can see from the figure that the model is relying on the timestamp in the upper left corner rather than the patient's figure or face, suggesting

it memorizes timestamps instead of distinguishing dementia. Additionally, backgrounds such as blackboards, calendars, and doors also influence classification due to different interview locations. Therefore, despite a high accuracy was achieved, the model's reliance on timestamps and backgrounds indicates institutional bias in the dataset, which makes the results unreliable.

Exp. II Classification of data instance processed by PaddleHub



Figure 4.8. An example of zooming in image data processed by the tool provided in the PaddleHub.

From the analysis above, we can see how the background could influence the classification of the machine learning models. Therefore, we believe it is important that we remove the influence of the background. We first tried the tool provided in the PaddleHub as stated in the proposal. PaddleHub could help cut out the background in most areas, however, as shown in Fig. 4.8, for that background that is embedded into deep sags or gaps between the hair strands, PaddleHub cannot properly cut them out. It also performs badly in what directly borders the shape edge of the human figure. The problem with this is that, even though we tried to erase most of the influencing factors in the pre-processing, the method itself cannot fully cut out the background. It often leaves a slim edge that is bound to the part of the human figure.

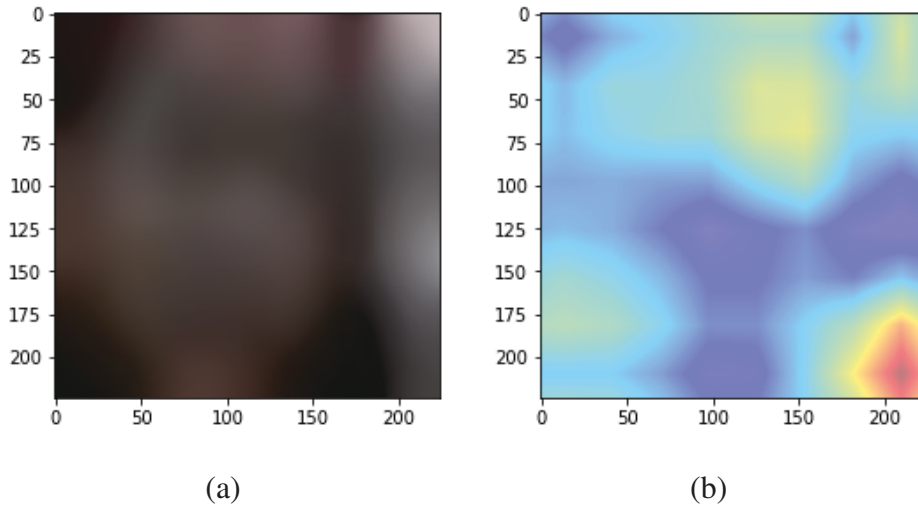


Figure 4.9. An example of image data cropped using PaddleHub and OpenCV with its corresponding heat map. (a) The face image that is blurred to protect the privacy of the patient. (b) The heat map corresponding to (a) that shows which part contributes to the classification.



Figure 4.10. An example of face mesh extracted from the image that contains a face using Media Pipe.

Exp. III classification of data instance processed by face mesh

As we checked the heat map (Fig. 4.9) of the processed images, we found that the convolutional network pays much attention to the edge of the face or the hair. After we carefully checked the processed image, we figured out that these areas usually contain some pixels of the background left out in the pre-processing of the data set just like Fig. 4.8. As

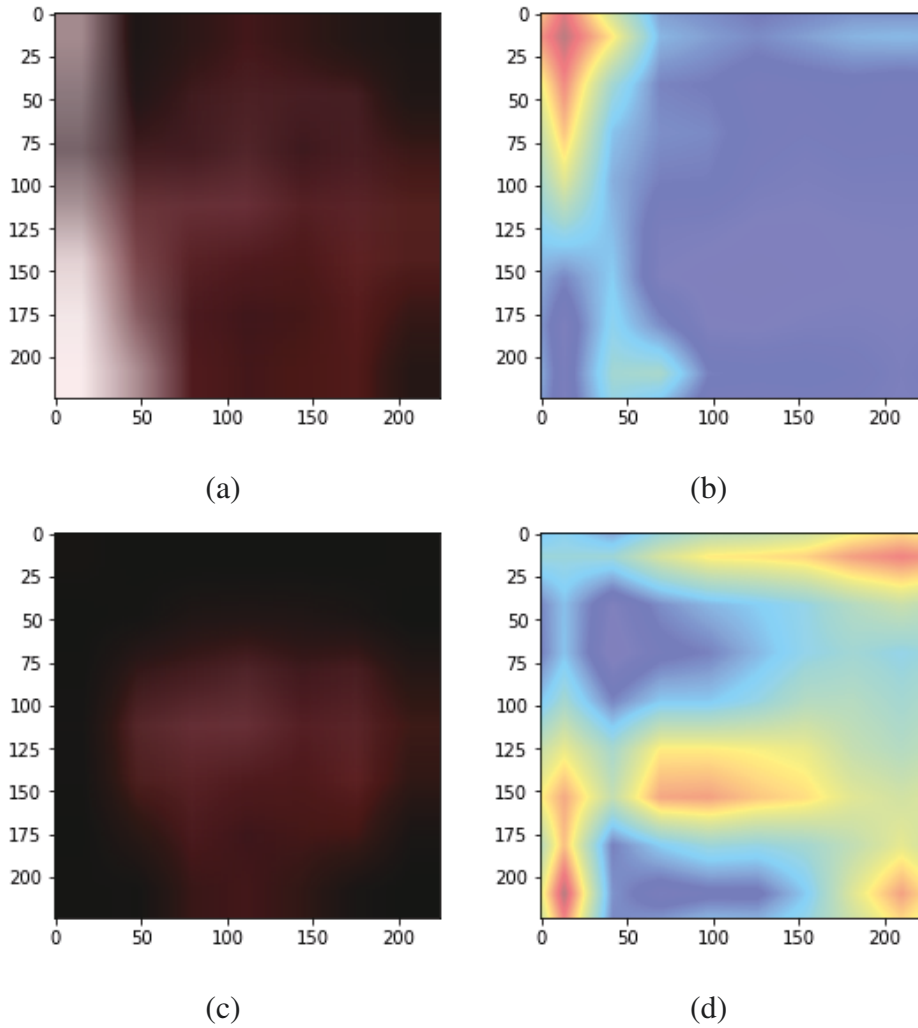


Figure 4.11. Examples of heat maps of the same picture but pre-processed with different tools. (a) the face image that is blurred and pre-processed by PaddleHub. (b) the corresponding heat map of (a). (c) the face image that is blurred and pre-processed using face mesh. (d) the corresponding heat map of (c).

CNNs are known to be sensitive to the change in the image, we suspected that these pixels are misleading the model. In this experiment, we achieved 79.65% of accuracy in the experiment after cropping the image with PaddleHub.

To solve the above-mentioned problems, we made use of face mesh landmarks that we extracted from the raw image data. An example of face mesh extracted using MediaPipe is shown in Fig. 4.10. We could notice in the example of the face mesh, it is composed of several parts including the eyes, nose, mouth, cheeks, and forehead. These landmarks are also made up of two contours that go around the shape of the face, with a bigger contour and a smaller one. In this research, to make sure that we fully cover the problems

Table 4.4. The accuracy performance when using different parts of the face to detect dementia.

	mouth	eyes	nose	cheek (control)
accuracy	77.20%	69.51%	75.55%	65.11%

mentioned in the last part, we just cut out the area circled by the smaller contour, which contains the sense organs and half of the forehead. We believe that, in this way, it would fully eliminate the influence of the background like the time stamp, the blackboard, or the calendar.

Fig. 4.11 shows the CAM heat maps of the same image but pre-processed with different tools. As shown in the figure, when we pre-process the image using PaddleHub, the model takes the upper left part of the image to be the main information source of why it thinks the image is from a dementia patient, while there may exist some pixels of the background that cannot be erased by PaddleHub. However, by using the face mesh to pre-process the image data, the focusing part of the model changes to other places in the image. Besides the focus on three corners, it also takes the nose and the mouth to be important. We achieved an accuracy of 83.06% in this experiment after cropping the image to the face mesh area.

Exp. IV Classification of data with individual sense organs

In the experiments for using sense organs to detect dementia, we achieved different accuracy on the validation set as shown in Table 4.4. As we can infer from the table, the mouth gives the best accuracy (77.20%) while the eyes get the least accuracy (69.51%) on the classification. The control area (cheek) also gives 65.11% of accuracy.

4.5.2 Classification Using 2D Facial Features

To compare the performance of our proposed method using 3D face mesh data against the method using simpler 2D facial features for dementia detection, we conducted a baseline experiment by implementing the method proposed by Tanaka et al. [64] on our datasets. we implemented the method proposed by Tanaka et al. [64] using the 2D facial features

Table 4.5. Top 20 Features by Absolute Coefficient in the Classification using 2D Facial Features.

Feature	Absolute Coefficient
AU12_r_mean	3.349433
AU10_r_mean	2.439014
AU14_r_mean	1.748242
AU45_r_std	1.647219
AU10_r_std	1.293248
AU26_r_std	1.133587
x_65_mean	0.907987
AU45_r_mean	0.887869
AU04_r_mean	0.825905
X_63_std	0.812892
AU17_r_std	0.808731
X_50_mean	0.786309
AU26_r_mean	0.761215
AU04_r_std	0.752868
X_48_mean	0.743195
y_60_mean	0.735887
x_50_mean	0.726954
x_63_std	0.725690
pose_Tx_std	0.718549
Y_61_mean	0.699022

extracted from the PROMPT dataset. We implemented the method using the same datasets and data split as we did for the face mesh data. This method using simpler 2D features yields an accuracy of 66.0%, with a precision of 60.1%, a recall of 62.7%, and an F1-score of 60.6%. Table 4.5 shows the top 20 features that contribute to the classification. The features are ranked by their absolute coefficients, which indicate the importance of each feature in the classification task. We can see that AU12_r_mean, AU10_r_mean, and AU14_r_mean are the top three features that contribute to the classification. Generally, the AU features are the most important features in the classification task, occupying 8

out of the top 10 features, which is consistent with the findings in [64]. However, the importance of individual AUs varies from the findings in [64], in which they reported that (in descending order) AU09, AU02, AU05, AU17, AU45, AU12, AU14, and AU04 are the most important AUs in the classification task.

4.5.3 Classification Using Face Mesh

Table 4.3 shows during the PCA reduction, how much variation in the original data can be explained with different numbers of components, which also indicates how much information was maintained after the dimensionality reduction was performed. The figure shows that 99% of the information can be retained and explained by 8 principal components after performing PCA on the face mesh data. PCA reduces the dimensionality of the data that needs to be processed, requires less computation, and helps relieve the over-fitting problem. We achieved an accuracy of 66% while using the normalized and dimensionality-reduced data. As a comparison, we achieved an accuracy of 59% using the normalized face mesh data without PCA dimensionality reduction. The experiments show that the face mesh data can potentially be employed in dementia detection. Besides, PCA reduction shows superior performance than its counterpart. However, even by reducing the dimension of the data, there was an over-fitting issue in the training curve, as the loss of the training data continued to decrease without improving the accuracy of the classification of the test set.

Besides for normalizing the face mesh data, we also tried to add more variations to the training data by inverting the normalization process. We have experimented with multiple neural networks including CNN, LSTM, and transformer.

Table 4.6 shows the performance of different deep learning models with different methods to add variations to the training data. The baseline model is the model trained on the normalized face mesh data without any additional variations. We can see the CNN model achieves the best accuracy of 75.9% when shifting is applied to the training data. Compared to the performance when normalization is applied, the accuracy of the improves by 10.6%. Therefore, our experiments show that adding variations to the training

Table 4.6. Dementia detection performance of deep learning models with different methods to add variations. Results are reported for accuracy, precision, and recall.

<i>Accuracy</i>	Models	Baseline	Rotation	Shifting	Scaling	All
	LSTM	0.698	0.705	0.705	0.707	0.636
	Transformer	0.730	0.748	0.752	0.725	0.676
	1D CNN	0.701	0.734	0.759	0.737	0.677
<i>Precision</i>	Models	Baseline	Rotation	Shifting	Scaling	All
	LSTM	0.719	0.722	0.716	0.792	0.619
	Transformer	0.765	0.710	0.709	0.691	0.696
	1D CNN	0.696	0.701	0.782	0.693	0.647
<i>Recall</i>	Models	Baseline	Rotation	Shifting	Scaling	All
	LSTM	0.512	0.533	0.543	0.451	0.444
	Transformer	0.553	0.717	0.738	0.676	0.464
	1D CNN	0.564	0.687	0.625	0.717	0.581

data can bring more benefits to the classification performance than just normalizing the data.

4.5.4 Classification Using HOG Features

Table 4.7. Classification performance in dementia detection using different numbers of frames of HOG features.

Number of Frames	Accuracy
1 frame	71.3%
8 frames	73.1%
1024 frames	79.0%

Table 4.7 shows the classification performance in dementia detection when different numbers of frames of HOG features were used. From the table, we can conclude that using only 1 frame of HOG features, the model can distinguish dementia patients with an accuracy of 71.3%. Using 8 frames of HOG features improves the accuracy to 73.1%. Using 1024 frames of HOG features improves the accuracy to 79.0%. The experiments show that HOG features can achieve a higher accuracy than face mesh or action units. However, it is worth noting that HOG features compute local gradients of pixels, which means that they do not only capture the edges of the face but also how the shades are distributed in the image. Therefore, HOG features are very likely to be influenced by the lighting conditions of the environment where the data was collected, which makes them unreliable for dementia detection. In our experiments, we achieved an accuracy of 71% when only 1 frame of HOG features was used, which is highly unlikely. This indicates that the HOG features are very likely to be influenced by the institutional bias, which makes them unreliable. Therefore, we cannot consider this performance as a reliable result in the current situation.

Table 4.8. Classification performance of the action units under different settings. Segment length refers to the window size when segmenting the video samples.

Segment Length / AU used	Max Accuracy	Recall	Precision	EER Accuracy
All frames & 11 AUs	69%	71%	66%	68%
512 frames & 11 AUs	70%	61%	73%	67%
1024 frames & 11 AUs	70%	66%	71%	67%
1024 frames & 17 AUs	71%	50%	85%	66%
1024 frames & 17 AUs	66%	42%	77%	63%

4.5.5 Classification Using Action Unit Features

Table 4.8 displays the classification performance of AUs in various scenarios. Our implemented voting system incorporates a variable parameter known as the threshold. This threshold represents a percentage value. During the testing phase, the tested subject has multiple video clips left out. These video clips are then classified using a trained SVM classifier. If the percentage of test video clips classified as healthy exceeds the threshold, we consider the test data to be healthy. Conversely, if the threshold is not exceeded, the subject is deemed to have dementia. In Table 4.8, the Max Accuracy column represents the optimal overall accuracy achieved by optimizing the threshold for best accuracy. Furthermore, the table provides the recall and precision values obtained when the threshold is optimized for the best accuracy. The EER Accuracy indicates the accuracy achieved by optimizing the threshold to ensure an equal error rate for dementia and healthy subjects. By segmenting the videos into 1024-frame clips, using all 17 AUs, we achieve the highest maximum accuracy of 71%. Alternatively, when we refrain from segmenting the videos and only employ 11 AUs that exhibit noticeable differences between healthy subjects and dementia patients, the best EER accuracy achieved is 68%.

Table 4.9 summarizes the performance of different methods in dementia detection. Despite the fact that our method using raw images achieves a high accuracy of 89%, it is biased by the lighting conditions and the environment where the data was collected. The method using HOG features achieves an accuracy of 79%, which is also biased by

Table 4.9. Performance comparison of different methods.

Method	Performance	Samples
Tanaka et al. [13]	0.82 (AUC)	33 samples
Umeda et al. [14] (raw images)	0.92 (Accuracy)	Biased results, 238 samples
Ours (raw images)	0.89 (Accuracy)	Biased results, 446 samples
Ours (2D features)	0.66 (Accuracy)	446 samples
Ours (HOG)	0.79 (Accuracy)	446 samples
Ours (Face mesh)	0.76 (Accuracy)	446 samples
Ours (Aus)	0.71 (Accuracy)	446 samples

the lighting conditions. The method using face mesh achieves an accuracy of 76%, which is more robust to the lighting conditions and the environment. The method using AUs achieves an accuracy of 71%, which is also robust to the lighting conditions and the environment. Our experiments show that raw images and HOG features are vulnerable to the bias caused by the lighting conditions and the environment where the data was collected. Face mesh and AUs are relatively more robust to the bias to our knowledge. Besides, as the face mesh data contains richer information than the raw images, it has the potential to distinguish diseases that cause similar syndromes. The face mesh data can also protect the patients’ privacy, as it does not contain any identifiable information about the patients. Moreover, the face mesh data can track the 3D movement of the patients, which can be manipulated and transformed for further processing. Therefore, we believe that face mesh and AUs are promising features for dementia detection and can be used in real-world applications.

4.5.6 Error Analysis

Due to the fact that the PROMPT dataset is collected from different institutions, we performed an error analysis to investigate whether the classification performance is affected by the institution where the data was collected. As in our previous experiments, we found that classification based on raw images and HOG features is vulnerable to the lighting conditions and the environment where the data was collected. Therefore, we performed

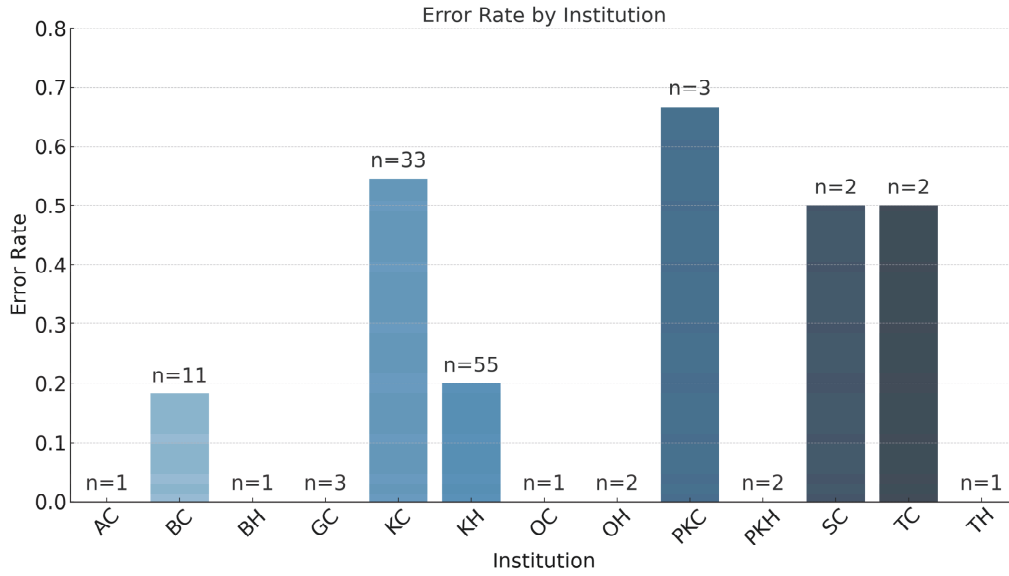


Figure 4.12. Error rate of the classification of dementia and healthy subjects using AUs. The x-axis shows the institution where the data was collected, and the y-axis shows the error rate of the classification.

an error analysis on the classification of dementia and healthy subjects using AUs. Figure 4.12 shows the error rate of the classification of dementia and healthy subjects using AUs. The overall accuracy is 0.70 with a p-value of 1.2×10^{-20} . The recall is 0.57, the precision is 0.74, and the F1 score is 0.65. The error rate across institutions is heterogeneous. We can see that the error rate of institutions KC and PKC is significantly higher than the error rate of other institutions. The institution of PKC has a small number of samples, which may lead to a higher error rate. However, the institution of KC has a large number of samples, which indicates that the error rate is reliable. This indicates that the classification performance is affected by the institution where the data was collected. Therefore, further data collection is needed to improve the classification performance and reliability of the results. Additionally, we also performed an error analysis on the classification of dementia and healthy subjects using face mesh. The overall accuracy is 0.78 with a p-value of 2.43×10^{-28} . The recall is 0.85, the precision is 0.78, and the F1 score is 0.81.

Figure 4.13 shows the error rate of the classification of dementia and healthy subjects using face mesh. The error rate across institutions is heterogeneous. However, compared to the error rate of the classification using AUs, the error rate for the institution of KC is

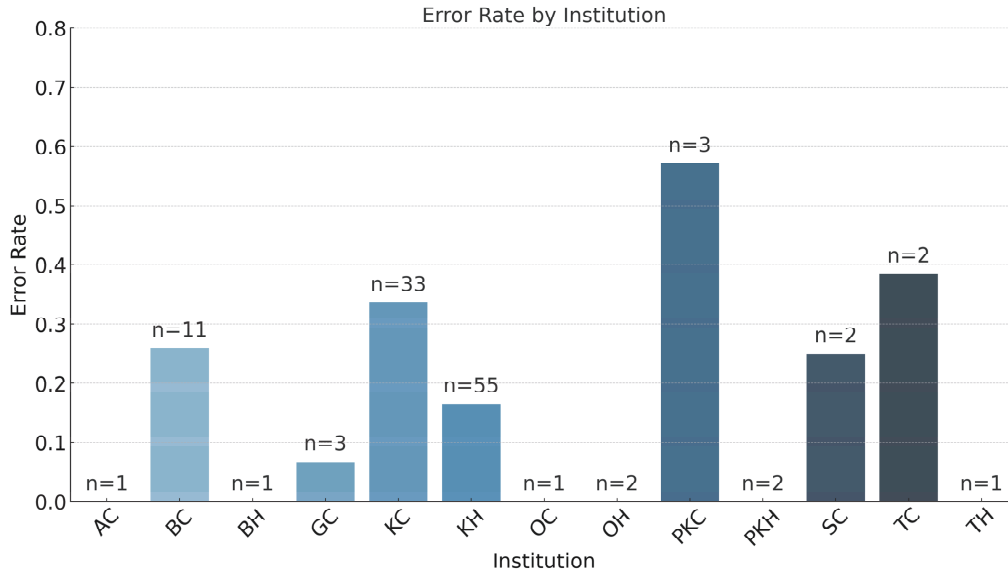


Figure 4.13. Error rate of the classification of dementia and healthy subjects using face mesh. The x-axis shows the institution where the data was collected, and the y-axis shows the error rate of the classification.

significantly lower, which indicates that the classification performance using face mesh is more robust to the institution where the data was collected.

In this section, we conducted an error analysis on the classification of dementia and healthy subjects using AUs and face mesh. The results show that the classification performance is affected by the institution where the data was collected. The error rate of the classification using AUs is heterogeneous across institutions, and the error rate of the classification using face mesh is more robust to the institution where the data was collected. This indicates that further data collection is needed to improve the classification performance and reliability of the results.

Despite we hope to perform a more detailed error analysis on the misclassified samples, we are facing the following challenges:

1. Low inter-annotator agreement among human annotators by directly observing the language and video data. Even the professional clinicians cannot tell if a person has dementia by just hearing the language or watching the video. In most cases, the diagnosis of dementia will combine clinical history, cognitive and neuropsychological assessments, physical and neurological examination, laboratory tests and brain imaging. These results will be considered comprehensively according to the diagnostic criteria as stated in the

DSM-5 [172]. The labels we have in the dataset are not annotated from the samples, either. They are decided by traditional methods and follow the criteria as in DSM-5. Therefore, by looking at the videos and listening to the audios, we cannot identify who is dementia or who is not (except for those who are severely demented).

2. Symptoms commonly associated with dementia, such as reduced speech fluency or diminished facial expression, are not exclusive to dementia and can also be observed in healthy individuals or those with other medical or psychiatric conditions. For example, reduced speech fluency may occur in people with depression, ADHD, or even in healthy individuals who are tired, stressed, distracted, or simply thinking. Similarly, normal aging is often accompanied by mild cognitive changes, such as occasional forgetfulness or word-finding difficulties, which do not necessarily indicate dementia. Diminished facial expression can also be seen in healthy individuals who are shy, introverted, or fatigued. Therefore, the presence of these symptoms alone is not sufficient for a definitive diagnosis of dementia, making it challenging to accurately interpret misclassified samples based solely on observable behaviors in language or video data.

3. Human reasoning in dementia detection is highly susceptible to bias from prior information. If we are told a subject has dementia, we are likely to interpret ambiguous behaviors (such as reduced speech fluency, stutters) as signs of dementia; if told the subject is healthy, we may reinterpret the exact same behaviors as normal. This cognitive bias makes it extremely difficult to objectively analyze misclassified samples, as our judgments are easily swayed by knowledge of the ground truth. Unlike tasks such as cat vs. dog image classification, where humans can confidently identify errors and attribute them to specific visual confounders (for example, “the model is classifying the cat into a dog because the model is confused by the fur color, the background, the occlusion”), dementia detection from language or video data lacks clear, observable markers that humans can reliably use for error analysis. As a result, even expert clinicians cannot consistently distinguish between healthy and dementia subjects based solely on these data modalities, making meaningful error analysis of misclassified samples particularly challenging.

4.6 Conclusion and Future Directions

In this chapter, we investigated the potential and performance of different facial features using different methods in dementia detection. In the experiment investigating the face mesh data, we found that 16 principal components can retain 99% of the information in the face mesh data. By performing the classification using the face mesh data with CNN model and adding shifting variations, we achieved an accuracy of 76%. In the experiment investigating HOG features, we achieved the best accuracy of 79%. However, according to the results, there may be a potential institutional bias due to the the vulnerability of HOG features to lighting conditions. In the experiment on AU data, we obtained an accuracy of 71% with proper segmentation, and we noticed that discarding some unimportant AUs does not worsen the performance. Our work shows that the facial features have the potential to work as an indicator in dementia detection, while protecting the privacy of the patients at the same time.

For future directions, we are interested in using explainable AI to explain the diagnosis of the classifier, which will promote the deployment of such detecting system in the real medical case. We aim to provide examinable clues to the doctors to aid them making further and detailed analysis on the situation of the patients.

Chapter 5

Conclusions and Future Work

The objective of this dissertation is to explore and develop novel methods for non-invasive, easy-to-use and affordable dementia detection using various modalities, including language and diagnosis videos. The research aims at addressing the 4 main challenges in dementia detection: data scarcity, lack of explanations, low data quality, and institutional bias. The contributions and conclusion of this dissertation are summarized below, along with suggestions for future work.

5.1 Contributions

Overall, this dissertation has made significant contributions to the field of dementia detection by exploring various modalities and developing novel methods based on machine learning to address the challenges of data scarcity, lack of explanations, low data quality, and institutional bias.

Specifically, in Chapter 2, we made significant contributions to the field of dementia detection by improving post hoc explanations and evaluations on the language model and perplexity-based methods. We proposed a novel method to separate the language components into syntactic and semantic components, and we systematically evaluated the contribution of each component to the classification performance. We showed that the syntactic component of the language is a strong indicator of dementia, and we demonstrated that explicitly embedding the syntactic component into the text can improve classification accuracy. Our findings suggest that both syntax and semantics are important for detecting dementia. Our methods provide a framework for improving the post hoc explanations by decomposing the language data, which allows for a precise analysis of the individual composition of languages.

In Chapter 3, we proposed a novel approach that leverages autoencoders to perform zero-shot dementia detection using healthy language data and binary classification using healthy and dementia language data. We systematically evaluated and visualized the performance of various autoencoder architectures and latent dimensions. The results indicated that the choice of embedding type has a more significant impact on classification

performance than the latent dimension size or architecture complexity. Our findings suggest that autoencoders can be a valuable tool for dementia detection, particularly when using advanced embeddings like LLaMA. Our approach addresses the challenge of data scarcity by enabling zero-shot detection using healthy language data only, thus reducing the need for large labeled dementia datasets, which is hard to collect. Additionally, our method also has the potential to achieve better post hoc explanations by employing the text embeddings from LLaMA, which can make self-explanations and provide insights into the classification reasoning.

In Chapter 4, we investigated the potential of using diagnosis video data in dementia detection. In this chapter, we first used explainable AI methods to analyze the video data and identify the potential institutional bias in the data by conducting a series of cloze test experiments. Our analysis revealed that the data is biased towards the institutional settings, such as the time stamp, the background, and the lighting conditions. Additionally, after addressing the bias, we explored the potential of using facial features such as face mesh, HOG features, and AUs in dementia detection. Specifically, for the action unit data, we proposed a novel method that combines segmentation and voting to improve the detection accuracy. We adapted and improved the method in [64] through segmentation and voting. We validated this method on a relatively large dataset, i.e., PROMPT dataset. For the face mesh data, we manually manipulated the 3-dimensional mesh to address the challenge of inconsistent postures and orientations in the video data. We investigated the effect of both normalization and adding more variations on the classification performance. The results showed that the facial features have the potential to work as an indicator in dementia detection, while protecting the privacy of the patients at the same time.

5.2 Conclusions

In this dissertation, we have explored various modalities for dementia detection, including language and diagnosis video data, and developed novel methods based on machine learning to address the challenges of data scarcity, lack of explanations, low data quality, and institutional bias. On the data scarcity challenge, we have proposed methods that

leverage language data and autoencoders to enable zero-shot detection using healthy language data only, thus reducing the need for large labeled dementia datasets. On the lack of explanation challenge, we have proposed methods that decompose the language data into syntactic and semantic components, allowing for a more detailed analysis of the individual composition of languages and providing justifiable and validatable explanations for the diagnosis. Besides, we have also proposed an effective system that uses autoencoders to reconstruct the text embeddings from large language models, which have the potential to make self-explanations and provide insights into the classification reasoning. To address the challenges of low data quality and institutional bias, we applied explainable AI techniques to analyze diagnosis video data and identify sources of bias by conducting a cloze test. We improved the data quality by developing normalization and processing pipelines to mitigate these biases and improve data consistency. Furthermore, we tackled the challenge of inconsistent postures and orientations in video data by manually editing the 3-dimensional face mesh, which successfully aligned the facial features across different samples and improved classification performance. We also explored the potential of using simpler facial features such as action units and HOG features in dementia detection, demonstrating that these features can be effective indicators while protecting patient privacy.

Among the methods explored, language demonstrated the strongest performance for dementia detection. Assessing language ability as a core symptom of dementia enables meaningful post hoc explanations and insights into underlying mechanisms, especially with recent advances in LLMs. Language data is also less privacy-sensitive than videos, making it a practical option for clinical use. Looking ahead, combining speech and facial features may further improve detection accuracy. These approaches are complementary: both speech and facial features can be extracted from interview videos, and modern smartphones facilitate simultaneous audio and video collection. Importantly, using facial features rather than raw video helps safeguard patient privacy, which is essential in healthcare applications.

Table 5.1. Comparison of traditional tests and speech- and facial expression–based detection methods for dementia detection.

Criterion	Traditional Tests (MRI / CT / Blood Tests)	Language-Based Detection	Facial Expression–Based Detection
Cost (Money)	High – equipment, staff, facility fees	Low – requires only a microphone	Low – requires only a camera
Cost (Time)	Medium to High – appointments, processing time	Low – minutes per session	Low – minutes per session
Usability	Low – requires trained personnel and clinical setup	High – easily deployed via apps	High – easily deployed via apps
Accessibility	Limited – urban hospitals or specialty clinics	High – remote or at-home	High – remote or at-home
Performance	High – clinically validated	Moderate – language is an important indicator	Moderate – “Apathy” cause expression changes
Privacy	Moderate – protected by medical data regulations	Risky / Secure – raw data / features	Risky / Secure – raw data / features
Invasiveness	Medium to High – includes blood draws, injection, radiation, or confined imaging	Low – passive and non-contact	Low – passive and non-contact

Table 5.1 summarizes the comparison between traditional tests and language- and diagnosis video-based detection methods for dementia detection. In this dissertation, we defined 7 key criteria: cost (money), cost (time), usability, accessibility, performance, privacy, and invasiveness:

- **Invasiveness:** How much a test or method physically affects the patient.
- **Cost (Money):** The financial expense of implementing the method, including equipment and facility requirements.
- **Cost (Time):** The amount of time needed for the test and result processing.
- **Usability:** How easy and efficient the method is to use. High usability methods require minimal training, while low usability methods may need specialized knowledge or equipment.
- **Accessibility:** How widely available the method is for patients. High accessibility methods can be used remotely or at home, while low accessibility methods require specialized clinics or equipment.

- Performance: The accuracy and reliability of the method in detecting dementia.
- Privacy: The protection of patient data during the diagnostic process. Secure privacy ensures data protection, while risky privacy may expose sensitive information.

Traditional tests, such as MRI, CT, and blood tests, are costly and time-consuming, requiring trained personnel and clinical setups. In contrast, language- and diagnosis video-based detection methods are low-cost, easy to use, and accessible, making them suitable for remote or at-home use. While traditional tests have high performance and are clinically validated, language- and diagnosis video-based methods have moderate performance, with language being an important indicator for dementia detection. These methods are better options for early detection and massive screening of dementia, especially in resource-limited settings.

Table 5.2. Summary of methods cited in Chapters 2, 3, and 4 with accuracy and evaluation remarks.

Chapter	Methods	Accuracy	Advantages / Disadvantages
Chapter 2	Wankerl et al. [98]	77%	Low computational cost / No post hoc explanation
	Fritsch et al. [101]	85%	Higher accuracy / No post hoc explanation
	Cohen et al. [102]	87%	High accuracy, explanation from vocabulary / Lack explanation from syntax
	Farzana et al. [89]	77%	Cross-dataset examination, reliable / Lack explanation, low reliability (used acoustic features)
	Karlekar [95]	91%	High accuracy / Imbalanced dataset chance = 80%, not reliable
	Ours	82%	Moderate accuracy, post hoc explanations on classification, reliable
Chapter 3	Li et al. [134]	76%	Employed GPT-2 but relatively low accuracy
	Runde et al. [135]	82%	Can make self-explanation, moderate accuracy, low data requirement / Require both labeled data
	Ours (zero-shot)	76%	Zero-shot detection, low data requirement, can make self-explanation / Accuracy lower than classification
	Ours (binary classification)	82%	Can make self-explanation, moderate accuracy, low data requirement / Require both labeled data
	Ours (by averaging each subject)	89%	Can make self-explanation, moderate accuracy, low data requirement / Require both labeled data
Chapter 4	Tanaka et al. [64]	82% (AUC)	Easy-to-implement methods / Lack reliability (33 samples only)
	Umeda et al. [63]	92%	Automatic feature extraction from raw images / Biased results
	Ours (raw images)	89%	Automatic feature extraction from raw images, identified institutional bias source / Biased results
	Ours (HOG)	79%	High accuracy in face-based detection / Likely to be biased results
	Ours (Face mesh)	76%	Reliable, high accuracy in face-based detection, high extension / High dimensionality and overfitting
	Ours (Aus)	71%	Reliable, high accuracy in face-based detection / Automatic AU detection accuracy needs improvement

Table 5.2 summarizes the performance of each method cited in Chapters 2, 3, and 4, along with their advantages and disadvantages. The table highlights the accuracy of each method and provides remarks on their usability and interpretability. Our methods generally achieve moderate to high accuracy, with the language-based methods showing advantage in making post hoc explanations and usability. The facial expression-based methods also demonstrate potential, especially with the use of face mesh and action units, although they face challenges related to data quality and institutional bias.

5.3 Applications and Future Work

We are currently planning to implement the language-based dementia detection methods into actual applications, particularly in social media platforms where users communicate through text or audio. By analyzing conversation logs, we can provide users with examinations of their language abilities and detect potential signs of dementia. The zero-shot method proposed in Chapter 3 allows us to identify anomalies in conversations and validate them through binary classification, making it a practical tool for early detection. For the video data-based dementia detection, we acknowledge that we are still in the early stages of research due to challenges such as institutional bias and low data quality. The precision of facial features is not yet sufficient for practical use without causing user annoyance. However, we are committed to addressing these challenges and improving the applicability of our methods in real-world settings.

The future work can be divided into three main directions: finer post hoc explanations, domain adversarial methods, and multimodal approaches. In the realm of finer post hoc explanations, future research can focus on developing methods to clarify the reasoning behind model predictions on the symptom and disease level, which can help healthcare professionals justify and validate their diagnoses. Explainability is crucial for building trust in AI systems, especially in healthcare applications where decisions can have significant implications for patient care. Recent years have seen the great improvement in large language models and their applications in various domains, including healthcare. Future

work can explore the integration of large language models with existing systems to generate more detailed and transparent explanations for clinical reasoning, patient-specific narratives, and actionable guidance. In the realm of domain adversarial methods, future research can focus on developing techniques to mitigate institutional bias in dementia detection models. This can involve creating domain-invariant representations that are robust to variations in data collection conditions, such as lighting, background, and camera angles. By addressing these biases, we can improve the generalizability and fairness of dementia detection models across different populations and settings. However, currently most of our data were collected in two institutions, where the environment and labels are highly entangled. Therefore, we need to collect more data from diverse institutions to ensure the robustness of the domain adversarial models. Domain adversarial networks can be a promising approach to address this challenge by learning domain-invariant features that are less sensitive to institutional biases. In the realm of multimodal approaches, future research can focus on integrating multiple modalities, such as speech, facial features, and other demographic features, to improve the accuracy and robustness of dementia detection. The combination of different modalities can provide a more comprehensive understanding of the patient's condition and enhance the overall performance of the detection models.

Chapter 6

Appendix

Appendix A

List of Author's Publications and Awards

A.1 Journals

1. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2022). An evaluation on information composition in dementia detection based on speech. *IEEE Access*, 10, 92294-92306.
2. Bouazizi, M., **Zheng, C.**, Yang, S., & Ohtsuki, T. (2023). Dementia detection from speech: What if language models are not the answer? *Information*, 15(1), 2.
3. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2023). Detecting dementia from face-related features with automated computational methods. *Bioengineering*, 10(7), 862.
4. Okunishi, T., **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2025). Dementia and MCI Detection Based on Comprehensive Facial Expression Analysis From Videos During Conversation. *IEEE Journal of Biomedical and Health Informatics*.

5. **Zheng, C.**, Bouazizi, M., Okunishi, T., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2025). Face Mesh for Dementia Detection: Evaluating Data Augmentation in Deep Learning and Traditional Machine Learning. *IEEE Access*, 13, 104501-104520.

A.2 Full Articles on International Conferences Proceedings

1. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2021, December). Dementia detection using two perplexity methods with part-of-speech tags. In 2021 International Conference on Emerging Technologies for Communications (ICETC). Tokyo, Japan.
2. Bouazizi, M., **Zheng, C.**, & Ohtsuki, T. (2022, January). Dementia detection using language models and transfer learning. In The 5th International Conference on Software Engineering and Information Management (ICSIM). Yokohama, Japan.
3. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2024). Dementia detection based on face image data and convolutional neural networks. In 2024 International Conference on Emerging Technologies for Communications (ICETC). Kitakyushu, Japan.
4. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2025, April). Enhancing dementia detection: Augmenting the face mesh. In IEEE International Symposium on Biomedical Imaging (ISBI 2025). Houston, USA.

A.3 Articles on Domestic Conference Proceedings

1. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2022, January). Dementia detection using two perplexity methods with part-of-speech tags. In IEICE Seminar for Sensor Networks and Mobile Intelligence. Nagano, Japan.

2. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2022, March). Dementia detection using two perplexity methods with part-of-speech tags. In IEICE General Conference. Online.
3. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2023, May). Dementia detection using the face mesh, action units, and HOG features. In IEICE Seminar for Sensor Networks and Mobile Intelligence. Okinawa, Japan.
4. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2024, January). Enhancing dementia detection: Exploring face mesh and deep learning approaches. In IEICE Seminar for Sensor Networks and Mobile Intelligence. Yamanashi, Japan.
5. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2025). Reveal Implicit Patterns with Large Language Models. In IEICE General Conference. Tokyo, Japan.

A.4 Presentations at domestic meetings

1. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2022, January). Dementia detection using two perplexity methods with part-of-speech tags. In IEICE Seminar for Sensor Networks and Mobile Intelligence. Nagano, Japan.
2. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2022, March). Dementia detection using two perplexity methods with part-of-speech tags. In IEICE General Conference. Online.
3. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2023, May). Dementia detection using the face mesh, action units, and HOG features. In IEICE Seminar for Sensor Networks and Mobile Intelligence. Okinawa, Japan.
4. **Zheng, C.**, Bouazizi, M., Ohtsuki, T., Kitazawa, M., Horigome, T., & Kishimoto, T. (2024, January). Enhancing dementia detection: Exploring face mesh and deep

learning approaches. In IEICE Seminar for Sensor Networks and Mobile Intelligence. Yamanashi, Japan.

5. **Zheng, C.**, Bouazizi, M., & Ohtsuki, T. (2025). Reveal Implicit Patterns with Large Language Models. In IEICE General Conference. Tokyo, Japan.

A.5 Awards

1. Young Researcher Encouragement Award. IEICE. Oct 13, 2023.
2. ICETC 2024 Best Student Award. IEICE. Nov. 26th, 2024.

References

- [1] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, “The diagnosis of dementia due to alzheimer’s disease: recommendations from the national institute on aging-alzheimer’s association workgroups on diagnostic guidelines for alzheimer’s disease,” *Alzheimer’s & dementia*, vol. 7, no. 3, pp. 263–269, 2011.
- [2] F. EDITION, “Diagnostic and statistical manual of mental disorders,” *American psychiatric association, Washington, DC*, pp. 205–224, 1980.
- [3] C. N. Harada, M. C. N. Love, and K. Triebel, “Normal cognitive aging,” *Clinics in geriatric medicine*, vol. 29, no. 4, p. 737, 2013.
- [4] S. K. Inouye, R. G. Westendorp, and J. S. Saczynski, “Delirium in elderly people,” *The lancet*, vol. 383, no. 9920, pp. 911–922, 2014.
- [5] F. Edition *et al.*, “Diagnostic and statistical manual of mental disorders,” *Am Psychiatric Assoc*, vol. 21, no. 21, pp. 591–643, 2013.
- [6] M. S. C. Almeida, L. F. d. Sousa Filho, P. M. Rabello, and B. M. Santiago, “International classification of diseases–11th revision: from design to implementation,” *Revista de Saúde Pública*, vol. 54, p. 104, 2020.
- [7] World Health Organization, “Dementia,” <https://www.who.int/news-room/fact-sheets/detail/dementia>, 2023, accessed: 2025-06-07.
- [8] Alzheimer’s Disease International, “Dementia statistics,” <https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/>, n.d., accessed: 2025-06-07.

- [9] M. M. Mielke, “Consideration of sex differences in the measurement and interpretation of alzheimer disease-related biofluid-based biomarkers,” *The journal of applied laboratory medicine*, vol. 5, no. 1, pp. 158–169, 2020.
- [10] M. Prince, M. Guerchet, and M. Prina, “The global impact of dementia 2013-2050,” 2013.
- [11] H. Amjad, D. L. Roth, O. C. Sheehan, C. G. Lyketsos, J. L. Wolff, and Q. M. Samus, “Underdiagnosis of dementia: an observational study of patterns in diagnosis and awareness in us older adults,” *Journal of general internal medicine*, vol. 33, pp. 1131–1138, 2018.
- [12] United Nations, Department of Economic and Social Affairs, Population Division, *World Population Prospects 2022: Summary of Results*. New York: United Nations, 2022. [Online]. Available: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf
- [13] E. Nichols, J. D. Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram *et al.*, “Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019,” *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022.
- [14] J. A. Bourgeois, J. Seaman, and M. E. Servis, “Delirium, dementia, and amnestic and other cognitive disorders,” *The American Psychiatric Publishing Textbook of Psychiatry*, p. 303, 2008.
- [15] A. Nandi, N. Counts, S. Chen, B. Seligman, D. Tortorice, D. Vigo, and D. E. Bloom, “Global and regional projections of the economic burden of alzheimer’s disease and related dementias from 2019 to 2050: A value of statistical life approach,” *EClinicalMedicine*, vol. 51, 2022.
- [16] A. Wimo, K. Seeher, R. Cataldi, E. Cyhlarova, J. L. Dielemann, O. Frisell, M. Guerchet, L. Jönsson, A. K. Malaha, E. Nichols, P. Pedroza, M. Prince,

- M. Knapp, and T. Dua, “The worldwide costs of dementia in 2019,” *Alzheimers. Dement.*, vol. 19, no. 7, pp. 2865–2873, Jul. 2023.
- [17] A. Andersson, L.-Å. Levin, and B. G. Emtinger, “The economic burden of informal care,” *International journal of technology assessment in health care*, vol. 18, no. 1, pp. 46–54, 2002.
- [18] W. H. Organization *et al.*, “Global status report on the public health response to dementia,” 2021.
- [19] P. Scheltens, K. Blennow, M. M. Breteler, B. De Strooper, G. B. Frisoni, S. Salloway, and W. M. Van der Flier, “Alzheimer’s disease,” *The Lancet*, vol. 388, no. 10043, pp. 505–517, 2016.
- [20] S. Tiwari, V. Atluri, A. Kaushik, A. Yndart, and M. Nair, “Alzheimer’s disease: pathogenesis, diagnostics, and therapeutics,” *International journal of nanomedicine*, pp. 5541–5554, 2019.
- [21] P. B. Gorelick, A. Scuteri, S. E. Black, C. DeCarli, S. M. Greenberg, C. Iadecola, L. J. Launer, S. Laurent, O. L. Lopez, D. Nyenhuis *et al.*, “Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the american heart association/american stroke association,” *stroke*, vol. 42, no. 9, pp. 2672–2713, 2011.
- [22] J. T O’Brien and A. Thomas, “Vascular dementia,” *The Lancet*, vol. 386, no. 10004, pp. 1698–1706, 2015.
- [23] I. G. McKeith, B. F. Boeve, D. W. Dickson, G. Halliday, J.-P. Taylor, D. Weintraub, D. Aarsland, J. Galvin, J. Attems, C. G. Ballard *et al.*, “Diagnosis and management of dementia with lewy bodies: Fourth consensus report of the dlb consortium,” *Neurology*, vol. 89, no. 1, pp. 88–100, 2017.
- [24] Z. Walker, K. L. Possin, B. F. Boeve, and D. Aarsland, “Lewy body dementias,” *The Lancet*, vol. 386, no. 10004, pp. 1683–1697, 2015.

- [25] K. Rascofsky, J. R. Hodges, D. Knopman, M. F. Mendez, J. H. Kramer, J. Neuhaus, J. C. Van Swieten, H. Seelaar, E. G. Dopper, C. U. Onyike *et al.*, “Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia,” *Brain*, vol. 134, no. 9, pp. 2456–2477, 2011.
- [26] J. Bang, S. Spina, and B. L. Miller, “Frontotemporal dementia,” *The Lancet*, vol. 386, no. 10004, pp. 1672–1682, 2015.
- [27] K. A. Jellinger, “The enigma of mixed dementia,” *Alzheimer’s & Dementia*, vol. 3, no. 1, pp. 40–53, 2007.
- [28] J. Gratwicke, M. Jahanshahi, and T. Foltynie, “Parkinson’s disease dementia: a neural networks perspective,” *Brain*, vol. 138, no. 6, pp. 1454–1476, 2015.
- [29] R. A. Roos, “Huntington’s disease: a clinical review,” *Orphanet journal of rare diseases*, vol. 5, pp. 1–8, 2010.
- [30] B. Sikorska, R. Knight, J. W. Ironside, and P. P. Liberski, “Creutzfeldt-jakob disease,” *Neurodegenerative Diseases*, pp. 76–90, 2012.
- [31] D. Shprecher, J. Schwalb, and R. Kurlan, “Normal pressure hydrocephalus: diagnosis and treatment,” *Current neurology and neuroscience reports*, vol. 8, no. 5, pp. 371–376, 2008.
- [32] J. Cerejeira, L. Lagarto, and E. B. Mukaetova-Ladinska, “Behavioral and psychological symptoms of dementia,” *Frontiers in neurology*, vol. 3, p. 73, 2012.
- [33] F. Müller-Spahn, “Behavioral disturbances in dementia,” *Dialogues in clinical neuroscience*, vol. 5, no. 1, pp. 49–59, 2003.
- [34] J. K. Harrison, P. Fearon, A. H. Noel-Storr, R. McShane, D. J. Stott, and T. J. Quinn, “Informant questionnaire on cognitive decline in the elderly (iqcode) for the diagnosis of dementia within a secondary care setting,” *Cochrane Database of Systematic Reviews*, no. 3, 2015.

- [35] J. Galvin, C. Roe, K. Powlishta, M. Coats, S. Muich, E. Grant, J. Miller, M. Storandt, and J. Morris, “The ad8: a brief informant interview to detect dementia,” *Neurology*, vol. 65, no. 4, pp. 559–564, 2005.
- [36] J. R. Cockrell and M. F. Folstein, “Mini-mental state examination,” *Principles and practice of geriatric psychiatry*, pp. 140–141, 2002.
- [37] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, “The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment,” *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.
- [38] J. Podhorna, T. Krahne, M. Shear, J. E. Harrison, and A. D. N. Initiative, “Alzheimer’s disease assessment scale–cognitive subscale variants in mild cognitive impairment and mild alzheimer’s disease: change over time and the effect of enrichment strategies,” *Alzheimer’s research & therapy*, vol. 8, pp. 1–13, 2016.
- [39] J. C. Morris, “Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the alzheimer type,” *International psychogeriatrics*, vol. 9, pp. 173–176, 1997.
- [40] J. Cummings, “The neuropsychiatric inventory: development and applications,” *Journal of geriatric psychiatry and neurology*, vol. 33, no. 2, pp. 73–84, 2020.
- [41] S. A. Greenberg, “The geriatric depression scale (gds),” *Best Practices in Nursing Care to Older Adults*, vol. 4, no. 1, pp. 1–2, 2012.
- [42] F. Massoud, G. Devi, J. T. Moroney, Y. Stern, A. Lawton, K. Bell, K. Marder, and R. Mayeux, “The role of routine laboratory studies and neuroimaging in the diagnosis of dementia: a clinicopathological study,” *Journal of the American Geriatrics Society*, vol. 48, no. 10, pp. 1204–1210, 2000.
- [43] P. Vemuri and C. R. Jack, “Role of structural mri in alzheimer’s disease,” *Alzheimer’s research & therapy*, vol. 2, pp. 1–10, 2010.

- [44] M. Pasi, A. Poggesi, and L. Pantoni, “The use of ct in dementia,” *International psychogeriatrics*, vol. 23, no. S2, pp. S6–S12, 2011.
- [45] C. Marcus, E. Mena, and R. M. Subramaniam, “Brain pet in the diagnosis of alzheimer’s disease,” *Clinical nuclear medicine*, vol. 39, no. 10, pp. e413–e426, 2014.
- [46] M. Chapleau, L. Iaccarino, D. Soleimani-Meigooni, and G. D. Rabinovici, “The role of amyloid pet in imaging neurodegenerative disorders: a review,” *Journal of Nuclear Medicine*, vol. 63, no. Supplement 1, pp. 13S–19S, 2022.
- [47] M. Schöll, A. Vrillon, T. Ikeuchi, F. Quevenco, L. Iaccarino, S. Vasileva-Metodiev, S. Burnham, J. Hendrix, S. Epelbaum, H. Zetterberg, and S. Palmqvist, “Cutting through the noise: A narrative review of alzheimer’s disease plasma biomarkers for routine clinical use,” *The Journal of Prevention of Alzheimer’s Disease*, vol. 12, no. 4, p. 100056, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2274580724007180>
- [48] H.-W. Klafki, M. Staufenbiel, J. Kornhuber, and J. Wiltfang, “Therapeutic approaches to alzheimer’s disease,” *Brain*, vol. 129, no. 11, pp. 2840–2855, 2006.
- [49] X. Guo, L. Gao, J. Liu, W. Lu, J. Wang, and Q. Qu, “Public acceptability of lumbar puncture in the diagnosis of alzheimer’s disease: a questionnaire-based single center study in china,” *American Journal of Translational Research*, vol. 15, no. 6, p. 4179, 2023.
- [50] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine *et al.*, “Toward defining the preclinical stages of alzheimer’s disease: Recommendations from the national institute on aging-alzheimer’s association workgroups on diagnostic guidelines for alzheimer’s disease,” *Alzheimer’s & dementia*, vol. 7, no. 3, pp. 280–292, 2011.
- [51] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

- [52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [55] P. K. Mall, P. K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, and M. Ganzha, “A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities,” *Healthcare Analytics*, vol. 4, p. 100216, 2023.
- [56] E. Altinkaya, K. Polat, and B. Barakli, “Detection of alzheimer’s disease and dementia states based on deep learning from mri images: a comprehensive review,” *Journal of the Institute of Electronics and Computer*, vol. 1, no. 1, pp. 39–53, 2020.
- [57] M. Perovnik, A. Vo, N. Nguyen, J. Jamšek, T. Rus, C. C. Tang, M. Trošt, and D. Eidelberg, “Automated differential diagnosis of dementia syndromes using fdg pet and machine learning,” *Frontiers in aging neuroscience*, vol. 14, p. 1005731, 2022.
- [58] M. R. Ahmed, Y. Zhang, Z. Feng, B. Lo, O. T. Inan, and H. Liao, “Neuroimaging and machine learning for dementia diagnosis: recent advancements and future prospects,” *IEEE reviews in biomedical engineering*, vol. 12, pp. 19–33, 2018.
- [59] D. Smirnova, T. Smirnova, and P. Cumming, “Language impairments in dementia: From word-finding difficulties to everyday conversation in a dementia-friendly

- community,” *Dementia Care: Issues, Responses and International Perspectives*, pp. 85–108, 2021.
- [60] K. H. Lee, J. Y. Lee, M. Boltz, and E. S. McConnell, “Emotional expression of persons with dementia: An integrative review with implications for evidence-based practice,” *Worldviews on Evidence-Based Nursing*, vol. 16, no. 5, pp. 344–351, 2019.
- [61] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, “What clinicians want: contextualizing explainable machine learning for clinical end use,” in *Machine learning for healthcare conference*. PMLR, 2019, pp. 359–380.
- [62] Y.-L. Liu, R. Feng, J.-H. Yuan, and Z.-H. Ling, “Clever hans effect found in automatic detection of alzheimer’s disease through speech,” *arXiv preprint arXiv:2406.07410*, 2024.
- [63] Y. Umeda-Kameyama, M. Kameyama, T. Tanaka, B.-K. Son, T. Kojima, M. Fukasawa, T. Iizuka, S. Ogawa, K. Iijima, and M. Akishita, “Screening of alzheimer’s disease by facial complexion using artificial intelligence,” *Aging (Albany NY)*, vol. 13, no. 2, p. 1765, 2021.
- [64] H. Tanaka, H. Adachi, H. Kazui, M. Ikeda, T. Kudo, and S. Nakamura, “Detecting dementia from face in human-agent interaction,” in *ICMI 2019*, 2019, pp. 1–6.
- [65] K. Ritchie, I. Carriere, L. Su, J. T. O’Brien, S. Lovestone, K. Wells, and C. W. Ritchie, “The midlife cognitive profiles of adults at high risk of late-onset alzheimer’s disease: The prevent study,” *Alzheimer’s & Dementia*, vol. 13, no. 10, pp. 1089–1097, 2017.
- [66] D. S. Geldmacher and P. J. Whitehouse, “Evaluation of dementia,” *New England Journal of Medicine*, vol. 335, no. 5, pp. 330–336, 1996.
- [67] H. C. Hendrie, “Epidemiology of dementia and alzheimer’s disease,” *The American Journal of Geriatric Psychiatry*, vol. 6, no. 2, pp. S3–S18, 1998.

- [68] J. O'brien, "Role of imaging techniques in the diagnosis of dementia," *The British journal of radiology*, vol. 80, no. special issue 2, pp. S71–S77, 2007.
- [69] K. Engedal, A. Brækhus, O. A. Andreassen, and P. H. Nakstad, "Diagnosis of dementia—automatic quantification of brain structures," *Tidsskrift for Den norske legeforening*, vol. 132, no. 15, pp. 1747–51, Aug. 2012.
- [70] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.
- [71] K. K. Zakzanis, S. J. Graham, and Z. Campbell, "A meta-analysis of structural and functional brain imaging in dementia of the alzheimer's type: a neuroimaging profile," *Neuropsychology review*, vol. 13, no. 1, pp. 1–18, 2003.
- [72] G. Bonifacio and G. Zamboni, "Brain imaging in dementia," *Postgraduate medical journal*, vol. 92, no. 1088, pp. 333–340, 2016.
- [73] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [74] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [75] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic alzheimer's disease detection from speech and language," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1784–1797, 2020.
- [76] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

- [77] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, “Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance,” *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.
- [78] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, “Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech,” in *IEEE International Conference Mechatronics and Automation*, 2005, vol. 3. IEEE, 2005, pp. 1569–1574.
- [79] F. Rudzicz, R. Wang, M. Begum, and A. Mihailidis, “Speech recognition in alzheimer’s disease with personal assistive robots,” in *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies*, 2014, pp. 20–28.
- [80] E. L. Campbell, L. Docío-Fernández, J. J. Raboso, and C. García-Mateo, “Alzheimer’s dementia detection from audio and text modalities,” *arXiv preprint arXiv:2008.04617*, 2020.
- [81] S. Luz, S. de la Fuente, and P. Albert, “A method for analysis of patient speech in dialogue for dementia detection,” *arXiv preprint arXiv:1811.09919*, 2018.
- [82] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, “Dementia detection using automatic analysis of conversations,” *Computer Speech & Language*, vol. 53, pp. 65–79, 2019.
- [83] A. Garcia-Silva, C. Berrio, and J. M. Gomez-Perez, “Understanding transformers for bot detection in twitter,” *arXiv preprint arXiv:2104.06182*, 2021.
- [84] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, “A comparative study of acoustic and linguistic features classification for alzheimer’s disease detection,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6423–6427.

- [85] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, “Machine learning for dementia prediction: a systematic review and future research directions,” *Journal of medical systems*, vol. 47, no. 1, p. 17, 2023.
- [86] G. Kantayeva, J. Lima, and A. I. Pereira, “Application of machine learning in dementia diagnosis: A systematic literature review,” *Heliyon*, vol. 9, no. 11, 2023.
- [87] Q.-T. Nguyen, L. Le, X.-T. Tran, T. Do, and C.-T. Lin, “Can chatgpt diagnose alzheimer’s disease?” *arXiv preprint arXiv:2502.06907*, 2025.
- [88] M. Shahin, B. Ahmed, and J. Epps, “Zero-shot cognitive impairment detection from speech using audiollm,” *arXiv preprint arXiv:2506.17351*, 2025.
- [89] S. Farzana, A. Deshpande, and N. Parde, “How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection,” in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 37–48.
- [90] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, “Spoken language derived measures for detecting mild cognitive impairment,” *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2081–2090, 2011.
- [91] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, “Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection,” vol. 2021, p. 3790, 2021.
- [92] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/>

- [93] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, “Aided diagnosis of dementia type through computer-based analysis of spontaneous speech,” in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.
- [94] C. Pope and B. H. Davis, “Finding a balance: The carolinas conversation collection,” 2011.
- [95] S. Karlekar, T. Niu, and M. Bansal, “Detecting linguistic characteristics of alzheimer’s dementia by interpreting neural models,” *arXiv preprint arXiv:1804.06440*, 2018.
- [96] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, “A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech,” *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.
- [97] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert *et al.*, “Automatic speech analysis for the assessment of patients with predementia and alzheimer’s disease,” *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.
- [98] S. Wankerl, E. Nöth, and S. Evert, “An n-gram based approach to the automatic diagnosis of alzheimer’s disease from spoken language.” in *Interspeech*, 2017, pp. 3162–3166.
- [99] D. Jurafsky and J. H. Martin, “An introduction to natural language processing, computational linguistics, and speech recognition,” 2000.
- [100] T. Dunning, “Statistical identification of languages-technical report mccs 94-273, algorithms-computer science,” *Computing Research Laboratory*, 1994.

- [101] J. Fritsch, S. Wankerl, and E. Nöth, “Automatic diagnosis of alzheimer’s disease using neural network language models,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5841–5845.
- [102] T. Cohen and S. Pakhomov, “A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the alzheimer’s type,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACM, 2020, pp. 1946–1957.
- [103] H. Bird, M. A. L. Ralph, K. Patterson, and J. R. Hodges, “The rise and fall of frequency and imageability: Noun and verb production in semantic dementia,” *Brain and language*, vol. 73, no. 1, pp. 17–49, 2000.
- [104] B. MacWhinney, “Talkbank,” *Recuperado de <http://talkbank.org>*, 1999.
- [105] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” *arXiv preprint arXiv:1801.06146*, 2018.
- [106] —, “Universal language model fine-tuning for text classification,” Jan. 2018, pp. 328–339.
- [107] S. Merity, N. S. Keskar, and R. Socher, “Regularizing and optimizing lstm language models,” *arXiv preprint arXiv:1708.02182*, 2017.
- [108] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, vol. 30, 2017, pp. 5998–6008.
- [109] P. Deneef and D. L. Kent, “Using treatment-tradeoff preferences to select diagnostic strategies: linking the roc curve to threshold analysis,” *Medical Decision Making*, vol. 13, no. 2, pp. 126–132, 1993.
- [110] A. Sommerlad, G. Perera, A. Singh-Manoux, G. Lewis, R. Stewart, and G. Livingston, “Accuracy of general hospital dementia diagnoses in england: sensitivity,

- specificity, and predictors of diagnostic accuracy 2008–2016,” *Alzheimer’s & Dementia*, vol. 14, no. 7, pp. 933–943, 2018.
- [111] E. Loper and S. Bird, “Nltk: The natural language toolkit,” *arXiv preprint cs/0205028*, 2002.
- [112] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, “A survey on large language models: Applications, challenges, limitations, and practical usage,” *Authorea Preprints*, 2023.
- [113] L. Floridi and M. Chiriatti, “Gpt-3: Its nature, scope, limits, and consequences,” *Minds and Machines*, vol. 30, pp. 681–694, 2020.
- [114] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [115] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, “The natural history of alzheimer’s disease: description of study cohort and accuracy of diagnosis,” *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.
- [116] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, “Deep learning for anomaly detection: A review,” *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.
- [117] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, “Deep semi-supervised anomaly detection,” in *International Conference on Learning Representations*.
- [118] J. Reilly, A. D. Rodriguez, M. Lamy, and J. Neils-Strunjas, “Cognition, language, and clinical pathological features of non-alzheimer’s dementias: an overview,” *Journal of communication disorders*, vol. 43, no. 5, pp. 438–452, 2010.
- [119] T. Cohen and S. Pakhomov, “A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the alzheimer’s type,” in

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1946–1957.

- [120] C. Zheng, M. Bouazizi, and T. Ohtsuki, “An evaluation on information composition in dementia detection based on speech,” *IEEE Access*, vol. 10, pp. 92 294–92 306, 2022.
- [121] P. Saltz, S. Y. Lin, S. C. Cheng, and D. Si, “Dementia detection using transformer-based deep learning and natural language processing models,” in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2021, pp. 509–510.
- [122] D. Ortiz-Perez, P. Ruiz-Ponce, D. Tomás, J. Garcia-Rodriguez, M. F. Vizcaya-Moreno, and M. Leo, “A deep learning-based multimodal architecture to predict signs of dementia,” *Neurocomputing*, vol. 548, p. 126413, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231223005362>
- [123] F. Bertini, D. Allevi, G. Lutero, L. Calzà, and D. Montesi, “An automatic alzheimer’s disease classifier based on spontaneous spoken english,” *Computer Speech & Language*, vol. 72, p. 101298, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230821000991>
- [124] Y. Santander-Cruz, S. Salazar-Colores, W. J. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, “Semantic feature extraction using sbert for dementia detection,” *Brain sciences*, vol. 12, no. 2, p. 270, 2022.
- [125] K. Lin and P. Y. Washington, “Multimodal deep learning for dementia classification using text and audio,” *Scientific Reports*, vol. 14, no. 1, p. 13887, 2024.
- [126] A. Baeveski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [127] T. Mikolov, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, vol. 3781, 2013.

- [128] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Detecting cognitive decline using speech only: The addresso challenge,” *arXiv preprint arXiv:2104.09356*, 2021.
- [129] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Zelasko, J. Villalba, and N. Dehak, “Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios.” vol. 2021, pp. 3825–3829, 2021.
- [130] F. Agbavor and H. Liang, “Artificial intelligence-enabled end-to-end detection and assessment of alzheimer’s disease using voice,” *Brain sciences*, vol. 13, no. 1, p. 28, 2022.
- [131] L. Ilias and D. Askounis, “Multimodal deep learning models for detecting dementia from speech and transcripts,” *Frontiers in Aging Neuroscience*, vol. 14, p. 830943, 2022.
- [132] M. Bouazizi, C. Zheng, S. Yang, and T. Ohtsuki, “Dementia detection from speech: what if language models are not the answer?” *Information*, vol. 15, no. 1, p. 2, 2023.
- [133] OpenAI, “Chatgpt: Large language model (january 2025 version),” 2025, accessed: 2025-01-25. [Online]. Available: <https://openai.com>
- [134] C. Li, D. Knopman, W. Xu, T. Cohen, and S. Pakhomov, “Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models,” *arXiv preprint arXiv:2203.13397*, 2022.
- [135] B. S. Runde, A. Alapati, and N. G. Bazan, “The optimization of a natural language processing approach for the automatic detection of alzheimer’s disease using gpt embeddings,” *Brain Sciences*, vol. 14, no. 3, 2024. [Online]. Available: <https://www.mdpi.com/2076-3425/14/3/211>

- [136] H. Torabi, S. L. Mirtaheri, and S. Greco, “Practical autoencoder based anomaly detection by using vector reconstruction error,” *Cybersecurity*, vol. 6, no. 1, p. 1, 2023.
- [137] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, “Autoencoders and their applications in machine learning: a survey,” *Artificial Intelligence Review*, vol. 57, no. 2, p. 28, 2024.
- [138] A. Ng *et al.*, “Sparse autoencoder,” *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.
- [139] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
- [140] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional autoencoders for hierarchical feature extraction,” in *Artificial neural networks and machine learning–ICANN 2011: 21st international conference on artificial neural networks, espoo, Finland, June 14-17, 2011, proceedings, part i 21*. Springer, 2011, pp. 52–59.
- [141] D. P. Kingma, M. Welling *et al.*, “Auto-encoding variational bayes,” 2013.
- [142] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, “Improving text embeddings with large language models,” *arXiv preprint arXiv:2401.00368*, 2023.
- [143] K. W. Church, “Word2vec,” *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.
- [144] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [145] R. Zeghari, A. König, R. Guerchouche, G. Sharma, J. Joshi, R. Fabre, P. Robert, V. Manera *et al.*, “Correlations between facial expressivity and apathy in elderly

- people with neurocognitive disorders: Exploratory study,” *JMIR formative research*, vol. 5, no. 3, p. e24727, 2021.
- [146] X. Xu, G. Jigeer, D. A. Gunn, Y. Liu, X. Chen, Y. Guo, Y. Li, X. Gu, Y. Ma, J. Wang *et al.*, “Facial aging, cognitive impairment, and dementia risk,” *Alzheimer’s Research & Therapy*, vol. 16, no. 1, p. 245, 2024.
- [147] C. Magai, C. Cohen, D. Gomberg, C. Malatesta, and C. Culver, “Emotional expression during mid- to late-stage dementia,” *International Psychogeriatrics*, vol. 8, no. 3, pp. 383–395, 1996. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S104161022400067X>
- [148] P. Ekman and W. V. Friesen, “Facial action coding system,” *Environmental Psychology & Nonverbal Behavior*, 1978.
- [149] P. Kaur, K. Krishan, S. K. Sharma, and T. Kanchan, “Facial-recognition algorithms: A literature review,” *Medicine, Science and the Law*, vol. 60, no. 2, pp. 131–139, 2020.
- [150] J. Weiner, M. Engelbart, and T. Schultz, “Manual and automatic transcriptions in dementia detection from speech.” in *Interspeech*, 2017, pp. 3117–3121.
- [151] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, “Detecting signs of dementia using word vector representations.” in *Interspeech*, 2018, pp. 1893–1897.
- [152] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, “Exploring deep transfer learning techniques for alzheimer’s dementia detection,” *Frontiers in computer science*, vol. 3, p. 624683, 2021.
- [153] P. Mahajan and V. Baths, “Acoustic and language based deep learning approaches for alzheimer’s dementia detection from spontaneous speech,” *Frontiers in Aging Neuroscience*, vol. 13, p. 623607, 2021.

- [154] D. Shibata, K. Ito, H. Nagai, T. Okahisa, A. Kinoshita, and E. Aramaki, “Idea density in japanese for the early detection of dementia based on narrative speech,” *PloS one*, vol. 13, no. 12, p. e0208418, 2018.
- [155] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech: The adress challenge,” *arXiv preprint arXiv:2004.06833*, 2020.
- [156] L. Ilias, D. Askounis, and J. Psarras, “Detecting dementia from speech and transcripts using transformers,” *Computer Speech & Language*, p. 101485, 2023.
- [157] K. Asplund, A. Norberg, R. Adolfsson, and H. M. Waxman, “Facial expressions in severely demented patients—a stimulus–response study of four patients with dementia of the alzheimer type,” *Int. J. Geriatr. Psychiatry*, vol. 6, no. 8, pp. 599–606, 1991.
- [158] K. Asplund, L. Jansson, and A. Norberg, “Facial expressions of patients with dementia: a comparison of two methods of interpretation,” *International Psychogeriatrics*, vol. 7, no. 4, pp. 527–534, 1995.
- [159] Y. Liu, Z. Wang, and G. Yu, “The effectiveness of facial expression recognition in detecting emotional responses to sound interventions in older adults with dementia,” *Frontiers in Psychology*, vol. 12, p. 707809, 2021.
- [160] Z. Jiang, S. Seyedi, R. U. Haque, A. L. Pongos, K. L. Vickers, C. M. Manzanares, J. J. Lah, A. I. Levey, and G. D. Clifford, “Automated analysis of facial emotions in subjects with cognitive impairment,” *Plos one*, vol. 17, no. 1, p. e0262527, 2022.
- [161] T. Kishimoto, A. Takamiya, K.-c. Liang, K. Funaki, T. Fujita, M. Kitazawa, M. Yoshimura, Y. Tazawa, T. Horigome, Y. Eguchi *et al.*, “The project for objective measures using computational psychiatry technology (prompt): Rationale, design, and methodology,” *Contemporary Clinical Trials Communications*, vol. 19, p. 100649, 2020.

- [162] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [163] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, “Mediapipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.
- [164] S. Abbasi and F. Tajeripour, “Detection of brain tumor in 3d mri images using local binary patterns and histogram orientation gradient,” *Neurocomputing*, vol. 219, pp. 526–535, 2017.
- [165] L. Ichim, D. Popescu, O. Chenaru *et al.*, “Advanced processing techniques for detection and classification of skin lesions,” in *2018 22nd International Conference on System Theory, Control and Computing (ICSTCC)*. IEEE, 2018, pp. 498–503.
- [166] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.
- [167] G. Bradski and A. Kaehler, “Opencv,” *Dr. Dobb’s journal of software tools*, vol. 3, p. 2, 2000.
- [168] Y. Ma, D. Yu, T. Wu, and H. Wang, “Paddlepaddle: An open-source deep learning platform from industrial practice,” *Frontiers of Data and Computing*, vol. 1, no. 1, pp. 105–115, 2019.
- [169] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *FG 2018*. IEEE, 2018, pp. 59–66.
- [170] T. Baltrušaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and person-specific normalisation for automatic action unit detection,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–6.

- [171] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “Constrained local neural fields for robust facial landmark detection in the wild,” in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 354–361.
- [172] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, “The dsm-5: Classification and criteria changes,” *World psychiatry*, vol. 12, no. 2, pp. 92–98, 2013.