A Multimodal Approach to Dementia Detection Using Contrastive Learning and LLM–VLM Assisted Reasoning with Guided Prompting

by

Fariha Zaman
24241328
Nowrin Sanjana
24241298

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

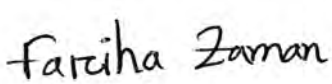
Department of Computer Science and Engineering
BRAC University
January 2026.

Declaration

It is hereby declared that

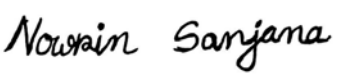
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Fariha Zaman

24241328



Nowrin Sanjana

24241298

Approval

The thesis/project titled "Advancing Dementia Prediction and Care Using Large Language Models (LLM's)"submitted by

1. Fariha Zaman (24241328)
2. Nowrin Sanjana (24241298)

of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Swakkhar Shatabda

Professor
Department of Computer Science and Engineering
Institution

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

The experiments ues data obtained from the DementiaBank pitt corpus and it contains anonymized speech recordings and transcripts collected under approved research protocols. No direct interaction with human participants was involved in this work.

All data were handled strictly for academic research purposes and no personally identifiable information was accessed or disclosed. The proposed system is intended as a research and decision support framework only and is not meant to replace professional medical diagnosis. Care was taken to ensure responsible use of CLAP Model, Large Language Models and Vision Language Models, with consideration of data privacy, transparency and potential model bias.

Abstract

Early detection of dementia remains challenging due to the cost, limited accessibility and subjectivity of traditional clinical assessments. This study proposes a non-invasive multimodal dementia detection framework using spontaneous speech from the Pitt Corpus of DementiaBank by exploring contrastive learning using text–audio representation and reasoning-based foundation models, Large Language Models and Vision Language Models. Linguistic features capturing lexical diversity, syntactic complexity, coherence,etc were combined with acoustic features including pitch, jitter, shimmer, MFCC,etc with Contrastive Language Audio Pretraining (CLAP) based text and audio embeddings. Among classical machine learning classifiers, the best performing configuration using Random Forest with text, audio and handcrafted features achieved an accuracy of 90.83%, F1-score of 0.9083 and AUC of 0.9478, while LightGBM achieved 89.91% accuracy, 0.8989 F1-score and the highest AUC of 0.9675, demonstrating the effectiveness of multimodal fusion of text, audio and features over unimodal baselines. In addition, large language models were evaluated under instruction based inference without fine tuning, where GPT-OSS achieved the highest accuracy of 68.07% among the tested LLMs, outperforming Qwen3-4B and Mistral. Vision language models were further examined using different prompting techniques and the hybrid VLM + LLM reasoning pipeline consistently outperformed standalone VLM configurations, indicating that hierarchical reasoning enhances multimodal dementia classification. Overall, the findings show that contrastive multimodal learning achieves strong classification performance, while reasoning based LLM and VLM frameworks enhance interpretability highlighting the potential of AI assisted methods for early dementia screening.

Keywords: Dementia Detection, Large Language Models (LLMs), Speech Analysis, Vision Language Models(VLMs), LLaVA-v1.6, Qwen2.5-VL, Multimodal, CLAP, Qwen, Mistral, GPT-OSS, Llava,SVM, Zeroshot, Chain of Though, Fewshot, Guided Prompt, Random Forest, XGBoost, LightGBM., SVM, KNN, LDA, Naive Bayes

Dedication

Our grandparents, who both bravely battled dementia, are honored in this thesis. Seeing what they experienced encouraged us to learn more about this condition and to conduct research that might help with early identification and awareness, even if in a small way. Their strength, patience and enduring impact on our lives remain the foundation of this work.

Acknowledgement

We begin by expressing our deepest gratitude to Almighty Allah for blessing us with good health, strength and the ability to accomplish our thesis work successfully and within the stipulated time. We are also thankful for the knowledge, skills and determination bestowed upon us, which enabled us to accomplish this task.

We would like to extend our sincere appreciation to our supervisor, Dr. Swakkhar Shatabda, Professor in Computer Science and Engineering, for his invaluable guidance, constant encouragement and constructive feedback throughout this research journey. His expertise and unwavering support have been pivotal in shaping our ideas and refining our work to its fullest potential.

The successful completion of this thesis would not have been possible without the collective efforts and support of many individuals. We are profoundly grateful to our parents, whose endless love, sacrifices and encouragement have been the driving force behind our academic pursuits. Their unwavering belief in us has made this achievement possible.

Lastly, we acknowledge the contributions of our peers, friends and all those who directly or indirectly supported us during this endeavor. Their encouragement and assistance have played a significant role in bringing this work to fruition.

Table of Contents

List of Figures

Chapter 1

Introduction

This section contains background, Rational of the Study or Motivation, Problem Statement, Objective, Methodology in Brief and Scopes and Challenges regarding our study.

1.1 Background

In accordance with the findings of "The 2021 Ageing Report" by the European Commission, the average lifespan has been steadily increasing over the course of the past several years. The number of individuals affected by dementia around the world is growing in tandem with the rise in the average lifespan [20]. Recent data of the World Health Organization (WHO) demonstrate the global dementia population is projected to reach 139 million in 2050 with an estimated 55 million people affected as of March 2023 and 10 million new cases diagnosed annually [45]. In a study, it is estimated that among 81.1 million dementia cases expected all around the world, 71% will take place in developing countries with yearly costs of $73 billion [41]. The disease can be caused by various medical conditions including traumatic brain injury (TBI), Huntington's disease, Down syndrome, Parkinson's disease and others. The diagnosis of dementia and the delay of its progression are both dependent on the early detection and intervention of the condition. In addition to the challenges posed by economical barriers and societal stigma, the difficulty of identifying symptoms from those of normal aging is another factor that contributes to the complexity of early detection [32]. Just 25% of those who are affected actually get a diagnosis [33]. These illnesses initiate silently up to 20 years prior to the manifestation of distinct cognitive signs. In the initial stage of dementia, participants exhibit changes in speech rhythm, characterized by an increased frequency of pauses, likely attributable to word-finding difficulties for example anomia and semantic paraphasia, diminished verbal fluency, reduction in the rate of speech and decreased spoken segments duration . These moderate impairments may stem from one or even many aspects at the level of words, sentences and discourse [20]. For example, patients with Alzheimer's disease often utilize semantically "empty" terms such as "thing" or "stuff" which lack definition and complexity. Moreover, they employ a reduced number of verbs that convey substantial information and a diminished quantity of nouns overall . Consequently, it may be more challenging to comprehend and believe their statements. Additionally, they may appear unorganized throughout their talk hindering their ability to engage in meaningful conversations [49].

A multitude of cognitive assessments such as the Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-Cognition (ADAS-Cog), Montreal Cognitive Assessment (MoCA) and Cognitive Dementia Rating (CDR) are currently employed to diagnose and track neurological conditions; however, their manual administration renders them expensive and often limits their accessibility to primary care. Among these, cookie theft description is a widely used test that evaluates both immediate and long-term recall, focus, the ability to concentrate and language skills, communication and organizational comprehension and can be administered fast but it is limited to clinical settings. Consequently, there is a growing need for rapid, noninvasive and automated digital biomarker delivery systems to facilitate real-world cognitive monitoring [20], [33].

Advances in digital technology including artificial intelligence (AI)-based systems are offering new opportunities for personalized dementia assessment and monitoring often complementing traditional methods like Mini-Mental State Examination (MMSE), the Alzheimer's Disease Assessment Scale-Cognition (ADAS-Cog) and Montreal Cognitive Assessment (MoCA) which often require costly tools, invasive procedures and time-consuming assessments [32]. Conventional Machine Learning (ML), Deep learning and Large Language Models (LLMs) can detect Alzheimer's Disease (AD) from linguistic, auditory and multimodal data. Machine learning models such as Support Vector Machines, XGBoost and Random Forest are employed to classify structured medical data particularly related to Dementia due to their integration of interpretability and predictive efficacy [30]. Language embeddings and structured data processing improve diagnostic accuracy in hybrid ML-LLaMA2-XGBoost models [51]. Machine learning excels at structured data tasks but deep learning models like BERT and GPT have improved speech and language-based dementia diagnosis by identifying patient spoken linguistic patterns [46]. Pause encoding for speech marker analysis [49] and multimodal categorization [45] are other deep learning methods. Deep learning is better at handling unstructured data than ML especially when fine-tuned on medical corpora [38]. LLMs use billions of parameters trained on enormous text data to capture intricate linguistic connections and synthesize human-like writing based on input prompts transforming medical diagnoses [33], [39]. GPT-3, GPT-4 and ChatGPT have been utilized for text embeddings in AD classification [20], zeroshot learning for diagnosis [44] and patient dialogue knowledge extraction [48]. Speech-based cognitive screening tools driven by ChatGPT and Bard [36] and LLM-assisted data annotation for clinical dataset labeling [50] have also been examined. Hybrid approaches combine LLM embeddings with ML-based classifiers and multimodal processing techniques to help LLMs handle complicated multimodal data [31]. Hybrid models that combine ML's structured decision-making with LLMs' contextual reasoning provide the most promise for dementia AD diagnosis [28], [37], [44].

1.2 Rational of the Study or Motivation

Dementia is a progressive brain disease that impacts cognition, language and memory. So, it is critical to discover the disease early in order to treat patients effectively.

Standardized assessments like the MMSE and ADAS-Cog are traditionally expensive and lengthy which makes them difficult to administer to some populations. In the field of dementia diagnosis, the transition from machine learning (ML) and deep learning (DL) to Large Language Models (LLMs) represents a substantial advancement in terms of both the capacities of models and the abilities to analyze data. Early machine learning models such as XGBoost and Random Forest performed exceptionally well in structured data tasks; however, they struggled when it came to unstructured data such as speech and text which are critical in the diagnosis of dementia. While deep learning methods like BERT-based audio classifiers have made great strides in speech analysis, there are still many obstacles to be conquered.The requirement for labels to be applied to large-scale datasets and the provision of processing resources is one of the problems that must be overcome. In contrast, deep learning algorithms have the ability to ignore the additional information that multimodal data might provide in favor of placing an excessive amount of focus on a single modality.This is because deep learning algorithms are structured to acquire knowledge. This makes their clinical use and their capacity to scale up more difficult. Deep learning and machine learning models both have intrinsic flaws including an over-reliance on labeled big datasets, a lack of contextual awareness and an incapacity to manage noisy or missing data.Through fewshot and zeroshot learning, LLMs and even multimodal VLMs could leverage their better natural language understanding and generative capabilities to improve contextual reasoning, extract high-level linguistic indicators and lessen dependency on labeled data addressing these shortcomings. The capability to comprehend intricate linguistic patterns has been significantly improved with the introduction of LLMs such as GPT-4 and LLaMA; however, these models do not yet fully incorporate speech or audio characteristics with their capabilities and mostly these llms are expensive. The need to overcome the limitations of existing machine and deep learning models using contrastive learning as well as utilize the potential of large language models and vision language models is the motivation this hard work. This study was motivated by significant gaps and limitations in the existing dementia detection methods which highlights the necessity of a solution that is both more efficient and more accessible.

1.3 Problem Statement

Significant gaps and limitations in the diagnosis of dementia related conditions are the driving force behind this research. The approaches which will be employed to achieve this goal: first, analyzing linguistic features and auditory patterns to find biomarkers using Contrastive Learning and second, testing how well Large Language Models (LLMs) and Vision Language models(VLMs) predict dementia severity.Through the utilization of various tools and libraries, this research aims to improve the detection rate of mild indicators of cognitive impairment through the evaluation of language and auditory biomarkers. Two dementia levels like Control and ProbableAD will be taken into account. Furthermore, it will assess how well LLMs and VLMs can use their advanced natural language processing (NLP) abilities.This approach not only improves our knowledge of the biomarkers linked to dementia, but it also offers a scalable solution that overcomes the flaws in both

previous machine learning deep learning technique and conventional diagnostic procedures. The motivation of this research is to ensure that patient care and outcomes are improved by early intervention. This will be achieved by facilitating early diagnosis that is both speedy and cost effective.

1.4 Objective

The objective of this study are:

- To analyze linguistic and acoustic biomarkers in order to predict dementia.
- To evaluate the effectiveness of CLAP-based multimodal representations for dementia classification and also exploring the contribution of linguistic and acoustic features.
- To investigate the impact of guided prompting strategies and varying model parameter scales on the performance of Large Language Models (LLMs) for dementia label detection without performing any model finetuning.
- To assess the performance of Vision Language Models (VLMs) under guided prompting examining their ability to reason over speech-derived textual and visual inputs and also a hybrid approach combining the strengths of both modalities: the VLM detects visual cues while the text-based LLM performs the final clinical reasoning.

1.5 Methodology in Brief

We use the The Pitt Corpus dataset by DementiaBank that includes the speech recorder and text transcript of people with Alzheimer and healthy individuals. Our method derives linguistic characteristics of a text including the use of pronouns, noun verb ratio, word richness, coherence, speech rate, speech duration, frequency of filler words, repetition, unfinished sentences, sentence length, syntactic complexity, depth of a parse tree, word content number and diversification of vocabulary with spaCy, NLTK, Scikit-learn and others.[9]. At the same time, such acoustic characteristics as pitch and voice stability (pitch, jitter, shimmer, HNR), formant frequencies and spectral features as the mean and the standard deviation of MFCC coefficients are calculated with Librosa.[10]. To enhance diagnostic accuracy, we want to develop a CLAP based multimodal model fusing linguistic and acoustic features. The model's performance is assessed using score of AUC besides accuracy, recall, F1-score and precision aiming for a more reliable and holistic approach to dementia prediction. We evaluate the capability of Large Language Models (LLMs) and Vision Language Models (VLMs) to forecast label of dementia via various methods of prompting in an instruction based inference task and do not finetune models. In the language only case, we use freely available LLMs, such as Qwen3-4B, Mistral-7B and GPT-OSS-20B. These models are used to create predictions based on the input data by using their instruction following capabilities. In the case of multimodal reasoning, we use VisionLanguage Models (VLMs) namely, Qwen-VL-Instruct and LLaVA-Vicuna-1.6-7B, which take both visual stimuli and textual descriptions to explain the level of

cognitive impairment. As well, we discuss a VLM + LLM hybrid model, in which the results of the VLMs are fed into Qwen-2.5-Instruct to improve cross-modal reasoning and decision making. All experiments are conducted without any parameter fine tuning or task specific training, relying solely on the models pre trained knowledge and instruction based prompting. Model performance is evaluated using standard classification metrics, including accuracy, precision, recall and F1-score, to assess their effectiveness in dementia level prediction.

1.6 Scopes and Challenges

This study will primarily focus on two modalities: speech and text transcripts for detecting dementia. Linguistic analysis will involve examining various aspects of speech, including syntactic complexity, lexical diversity, coherence and fluency which are known to be affected by cognitive decline. Alterations to vocal traits, such as pitch, rhythm, volume and pauses can be an indicator of neurological decline which is why acoustic analysis is so important. In order to assess the severity of dementia, these two methods will be used as a basis.

A key challenge of this study is the reliance on small, free and non–fine-tuned LLMs and VLMs, which may have limited reasoning capacity and domain specialization compared to large proprietary models. Such models might not be able to represent fine linguistic and acoustic signals related to early stage cognitive impairment. Also, age, education and recording conditions vary speech and can provide noise that can influence feature-based analysis and model driven analysis. Acoustic characteristics are prone to recording quality and environmental noise. Limited sample size and the challenge of separating changes in dementia related speech with other neurological or psychiatric disorders also make quality assessment difficult.

Chapter 2

Literature Review

This section includes an overview of Machine Learning Algorithms, Deep Learning Algorithms, Large Language Models and Prompting Techniques, followed by a review of existing research and concludes with a summary of key findings, including key results, patterns, trends and scope for improvements.

2.1 Preliminaries

This section is organized into subsections that provide a thorough understanding of fundamental ideas in machine learning (ML), deep learning (DL), large language models (LLMs) and prompting strategies. It begins with basic ML techniques like Support Vector Machine (SVM), Logistic Regression (LR), Ridge Regression and others. The discussion then shifts to advanced Deep Learning architectures such as BiLSTM with Attention, BERT, RoBERTa and ALBERT. The part also investigates cutting edge LLMs such as GPT-4, LLaMA-2, PaLM and SLIME. Finally, it covers prompting techniques such as FewShot Learning and Chain-of-Thought Reasoning. The section aims to build a clear understanding of these technologies, their applications, limitations and their relevance in solving real-world problems.

2.1.1 Machine Learning Algorithms:

Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised machine learning method used in machine learning categorization problems. In a N-dimensional space when N denotes feature number, SVM finds a hyperplane to clearly separate data points. The best hyperplane maximizes gap between plane and nearby data points of each class known as support vectors. SVM delivers strong results for high-dimensional spaces handling overfitting especially when there's a clear distinction between classes. Highly effective in high-dimensional spaces as it can use different kernel functions like linear, polynomial, RBF, etc. Although it works well with small datasets, it finds difficulty dealing with noisy datasets and computationally expensive for very large datasets [2].

Logistic Regression (LR):
Logistic Regression (LR) is a statistical model used for binary classification activities which is computationally fast, easy to understand and performs well when the data can be separated with a straight line. Logistic Regression predicts probabilities and

with the help of the sigmoid function, maps them to distinct classes unlike linear regression . The model finds the suitable parameters by keeping the probability of the observed data optimal. Logistic Regression is a simple yet powerful baseline model widely used in fields like medical diagnosis, spam detection, etc. It's performance goes down with nonlinear problems and when predictors are highly correlated) it assumes that the predictors are independent of each other [6].

Random Forest:
Random Forest is an ensemble learning algorithm that functions by creating multiple decision trees during training process. For classification tasks, it selects the class that's approved by the majority of trees. When it comes to regression tasks, it calculates the average of the predictions. The model operates feature selection and data sampling in a random process making it more robust and thus overfitting is reduced. It is well-suited for structured data and handles large datasets even with imbalanced and missing data but slower for large datasets and is harder to interpret compared to a single decision tree [4].

Ridge Regression:
Ridge Regression is an advanced type of linear regression that addresses prevalent issues such as multicollinearity, where input characteristics exhibit strong intercorrelation and overfitting, which arises when a model is too customized to the training data, resulting in subpar performance on fresh data. In contrast to traditional linear regression, Ridge Regression integrates a penalty word into the models loss function. Simplifying the model and minimizing excessive complexity decreases the influence of less important elements. To establish an optimal balance between a simple model and good data fitting, Ridge Regression is employed to alter the severity of the penalty. This maintains the model's durability and stability, which is especially useful when managing datasets with many related features. idge Regression is a flexible method that can be used to model financial patterns, such as stock returns predict events, such as real estate values and study complex biological datasets with related characteristics. It also takes size, age and location into account. Evaluating complex biological datasets with connected characteristics, predicting financial patterns, including stock returns and examining the impact of factors such as age, geographic location and size on real estate prices. By reducing the impact of less important features, managing datasets with linked variables and providing user friendliness, it is excellent at reducing overfitting. On the other hand, it keeps all features unfiltered and could perform less well if the regularization is particularly powerful [1].

Random Forest Regression (RFR):
Random Forest Regression is a consistent and robust machine learning algorithm. It involves developing multiple decision trees and connecting their predictions in order to improve accuracy and reduce overfitting. At each split, a random subset of attributes is employed and each tree is trained on a randomly chosen portion of the data. This guarantees that there is diversity within the trees and that the model operates effectively on the data collected. RFR prevents overfitting, manages non linear interactions and generates correct results by averaging the predictions of all trees. It provides insights of the most important features by making it an ideal tool for data comprehension. RFR is used in many different domains, including agricul-

tural production estimation, projections of sales and customer retention forecasting. Compared to simpler models, it can be difficult to understand. Large datasets can make it slow and in order to get the best results, parameters like tree depth and number of trees must be carefully adjusted [4].

Support Vector Regression (SVR):
Support Vector Regression (SVR) is a version on the Support Vector Machine (SVM) technology that was designed specifically for regression problems. Rather than fitting every data point precisely, SVR accommodates for outliers and noisy data by allowing a small margin of error known as epsilon. It is also capable of processing complex data with patterns and non linear correlations when kernel functions such as polynomial kernels or radial basis functions (RBF) are applied. Support vector regression (SVR) is commonly used in time-series forecasting (for variables such as energy consumption or weather, healthcare outcome prediction (for variables such as recovery time) and financial trend analysis variables like as stock prices. However, compared to linear regression and other simpler models, it requires fine tuning of parameters like the kernel and epsilon, which is less interpretable and can be slow with large datasets. Despite the drawbacks, its adaptability and capacity to manage high dimensional data make it a priceless tool for several applications [5].

Extreme Gradient Boosting (XGBoost):
XGBoost is considered to be an enhanced version of gradient boosting. It is a machine learning technique which builds models one after one where each new model aims to correct the mistakes of the previous one. In order to make the process efficient and faster, XGBoost enhances the process by introducing techniques like parallelization, regularization and tree pruning. It is a common choice in data science competitions as it is highly effective for tabular data. Handles missing values very well and includes built-in regularization to prevent overfitting but needs careful hyperparameters tuning and if not properly regulated, it can induce overfitting [8].

2.1.2 Deep Learning Algorithms:

Bidirectional Long Short-Term Memory (BiLSTM):
BiLSTM is a recurrent neural network (RNN) architecture designed to process sequential data by considering information from both directions (past to future and future to past). Unlike unidirectional LSTMs which only take past information into account, BiLSTMs make it especially effective for tasks where both previous and future contexts are essential. It is widely used in sentiment analysis, named entity recognition (NER), machine translation, speech recognition, etc. Although they capture bidirectional dependencies resulting in better performance on sequence modeling tasks but they are computationally more demanding than unidirectional LSTMs [3].

BiLSTM with Attention:
BiLSTM with Attention is built on the BiLSTM architecture by adding an attention mechanism. This attention mechanism dynamically assigns weights of importance to different parts of the input sequence so the model can focus on the most relevant tokens while making predictions. This is very useful for tasks involving long sequences

like machine translation text summarization, question answering and speech-to-text where certain words in the input are more crucial for producing accurate translations and as a result improves performance. The models complexity makes it more computationally demanding than traditional models and also requires precise tuning to maintain performance particularly when working with noisy data [7].

Bidirectional Encoder Representations from Transformers (BERT):
BERT is a transformer based model designed for natural language understanding tasks analyzing words in context from both directions like left and right of a word in a sentence. As it is pretrained on vast datasets, BERT can be fine-tuned for specific tasks like text classification, NER and question answering. It provides state-of-the-art performance on a variety of NLP benchmarks but computational cost for training is high and fine-tuning is resource-intensive [11].

Robustly Optimized BERT Approach (RoBERTa):
RoBERTa, an optimized version of BERT, improves its predecessor's performance by extending the pretraining process. By removing the next sentence prediction task , it increases the training data and batch sizes and trains on longer sequences with dynamic masking patterns leading to improved language modeling capabilities. It makes sure of improved accuracy over BERT on most NLP tasks like text classification and summarization. RoBERTa needs significant computational resources and large datasets which may not be accessible for many individuals with limited computing power. Despite it shows better improvements over BERT, it faces difficulties with very large datasets [12].

A Lite BERT (ALBERT):
As a lightweight version of BERT, ALBERT is designed to reduce model size and training time while improving performance by techniques like factorized embeddings and cross-layer parameter sharing. ALBERT offers outstanding performance with fewer parameters, making it suitable for tasks such as text classification, question answering, etc. It implements lower memory usage and faster training compared to BERT but sometimes struggles with considerable resources for optimal results [17].

DistilBERT:
DistilBERT, a more efficient variant of BERT, is a smaller and faster version of BERT created through knowledge distillation. It maintains 97% of BERT's language understanding capabilities without significantly compromising performance while being 40% smaller and 60% faster. Highly suitable for applications where computational resources are limited offering quicker processing for real-time applications. Although DistilBERT is cost-effective and requires less computational power to train, it may struggle with complex tasks, show more biases and its performance depends on the quality and quantity of fine-tuning data [13].

BioBERT:
BioBERT is a BERT based model fine-tuned on large biomedical datasets making efficient for tasks like biomedical named entity recognition (NER) and relation extraction. It outperforms BERT in biomedical-specific tasks by accurately identifying biomedical entities such as genes, proteins, diseases and extracting relationships be-

tween them. Although BioBERT is highly effective in biomedical applications but it requires significant computational resources especially when fine-tuning on large datasets. While its "black box" nature makes it hard to understand, its performance is dependent on the quality of the fine-tuning that can impact the accuracy of its results [18].

Wav2Vec:
Wav2Vec is a self-supervised learning model that can process raw audio to create speech representations using the help of a convolutional neural network and Transformer. It captures features from speech in a way that makes it highly effective for downstream tasks like automatic speech recognition (ASR), audio classification and speaker identification. Wav2Vec 2.0 introduced a more robust training method by combining contrastive learning with transformer models. As it removes the need for manual feature engineering in speech tasks, multilingual applications can be executed. The model works well with low-resource and limited labeled data by leveraging large amounts of unlabeled speech during pre-training. But it requires substantial computational power and careful fine-tuning to deliver the best results for specific tasks [15].

2.1.3 Large Language Model:

Generative Pre-training Transformer 3 (GPT-3):
At the time when GPT-3 launched, GPT-3 is considered to be the most capable one among the language models available then. It is capable of performing a great number of different activities the creation of text, translating, summarizing and answering questions. It does this by providing the transformer architecture that helps it produce text that is semantically consistent as well as contextually pertinent. The neural network of the model was trained on a huge and varied dataset that was extracted by books, articles and the internet. This capacity led to the acquisition of the ability to generalize successfully on a large scale of applications. The advantage of GPT-3 is that, it can work in an environment but this environment is a fewshot or zeroshot environment. It implies that one will be able to perform the work with the minimum or no task specific examples at all. GPT-3 network is a Large Language Model (LLM) that made a revolution in natural language processing (NLP) by showing that as a model gets larger and the volume of data in it increases, new abilities are unlocked. It was a by-product of this that it became a standard to the research and implementation of current LLM [16].

Generative Pre-training Transformer 4 (GPT-4):
OpenAI launched GPT-4 in 2023 and it is the subsequent evolution in the GPT series and has multimodal functionalities, which aloows it to handle both text and image inputs. This improvement allows GPT-4 to accomplish tasks that require both written and visual understanding like those related to understanding graphs or analyzing image captions. It is built on the transformer architecture and it incorporates improvements in reasoning and alignment with user intent with techniques like Reinforcement Learning with Human Feedback (RLHF). In addition, it outperforms previous versions in ambiguous questions, complicated problem solving and

delivering more accurate and factually correct replies. GPT-4 depicts the growth of LLMs from text only systems to adaptive, multimodal models, broadening their applications in domains such as education, healthcare and AI enhanced creativity. Its multimodal qualities represent the forefront of LLM research, expanding the limitations of AI's interaction with the world [26].

Large Language Model Meta AI 2 (LLaMA-2):
Large Language Model Meta AI (LLaMA-2) is free and therefore is an important resource to use in research and development based activities. The model uses the transformer architecture and is trained on a different form of dataset. It enables it to perform many NLP jobs, such as sentiment detection, text summarization and question answering. Compared to the bigger models, it pays more attention to the efficiency of parameters since it is able to gain competitive performance. LLaMA-2 is an important step towards enabling more researchers to access LLaMA technology and it opens opportunities to study the progress in natural language processing (NLP) without any limitations imposed by restricted access. The open source characteristic highlights the significance of transparency and collaboration in enhancing the functionalities of LLMs [29].

LLaMA 3.2-3B-Instruct:
LLaMA 3.2 3B Instruct, a component of Meta's 2024 LLaMA series is a succinct model with 3 billion parameters, tailored for instruction-following tasks like as answering queries, summarizing content and composing text. Despite its compact dimensions, it provides robust performance by superior training, meticulous fine-tuning and architectural enhancements. Engineered for efficiency, it operates effectively on consumer GPUs and mobile devices. Distributed under a permissive license, it is available for both research and commercial applications, facilitating seamless integration through platforms such as Hugging Face [42].

Pathways Language Model (PaLM):
The Pathways Language Model (PaLM) is a collection of extensive transformer based language models that were developed by Google AI. These models were designed for natural language processing (NLP) operations such as text generation, code completion and reasoning. PaLM makes use of the Pathways system, which makes it possible to conduct effective training over a large number of TPUs (Tensor Processing Units). This also makes it possible to scale effectively while maintaining excellent performance. In a manner that is comparable to that of GPT models, the PaLM design makes use of dense decoder-only transformers. However, it is strengthened by greater data efficiency and training methodologies. To improve its reasoning and linguistic capabilities, it makes use of techniques like sparsity, parallelism and pre-training on a diverse corpus. PaLM outperforms earlier models in tasks such as common-sense reasoning and arithmetic problem solving, demonstrating performance that is nearly identical to that of humans in certain domains [21].

Structured Language-Informed Multitask Embeddings (SLIME):
Structured Language-Informed Multitask Embeddings (SLIME) is a framework which is developed to improve the efficiancy of NLP models. It is improved through the use of structured data and language-informed embeddings for multitask learning. In

contrast to conventional explainability methods such as LIME (Local Interpretable Model-agnostic Explanations), SLIME use self-supervised learning to produce more dependable and significant feature attributions. It utilizes pre-trained language models, such as BERT or GPT to produce embeddings that encompass both linguistic context and structured information relevent to a specific job. SLIME improves models understanding of the connections between data features and language by merging structural and textual representations, hence enhancing performance in knowledge retrieval, recommendation systems and classification. This model is particularly advantageous in situations where structured data, such as tables or metadata must augment unstructured textual data to yield more accurate predictions. SLIME consistently incorporates multitask training, task-specific objectives and contextual embeddings to ensure the model's effective adaptation to diverse tasks [19].

Qwen 3 (7B):
Qwen 3 7B is developed via Alibaba Cloud in 2024. It is an advanced large language model utilizing the transformer architecture. Having been trained on more than 3 trillion tokens it demonstrates proficiency in tasks such as conversation, reasoning and language comprehension. Significantly, it exhibits strong performance even prior to fine-tuning and its chat-optimized variant (Qwen 3 7B-Chat) competes with larger models such as GPT-3.5. Utilizing sophisticated methods such as grouped-query attention and rotary positional embeddings (RoPE), Qwen 3 is both rapid and contextually aware. Disseminated under a commercially advantageous open license, it exemplifies Alibaba's dedication to providing accessible, high-performance AI technologies [52].

Gemma 2B IT:
In 2024, Google DeepMind unveiled Gemma 2B IT. It is a 2 billion parameter, lightweight, instruction-tuned model optimized for local devices. Based on the transformer base of Gemini and PaLM 2 it has been fine-tuned with RLHF to perform very well in tasks like communication, reasoning and summarization. Despite its modest size, it supports popular machine learning frameworks like PyTorch, TensorFlow and JAX and works well in low-resource environments. Gemma strikes a mix between developer accessibility and usability and is released with open weights under a responsible-use license [34].

Phi-3 Mini (4K) Instruct:
Microsoft Research unveiled the small 3.8-billion-parameter Phi-3 Mini 4K Instruct model in 2024. It is engineered for efficiency and strong instruction-following. Longer chats and larger papers are no problem for its 4,000-token context frame. It outperforms bigger models, thanks to its training on high-quality filtered and generated data. The algorithm has been fine-tuned for safe and useful responses, making it ideal for real-world application on mobile and low-power devices. Included in the release is documentation and pre-trained weights, both of which are available under an open license[43].

2.1.4 Prompting Techniques:

Zero Shot:
Zeroshot prompting is a technique where the model is given a task without any prior examples or specific training for that task. The prompt provides direct instructions or questions and the model generates a response based purely on its pre-trained knowledge. This approach leverages the models ability to generalize across tasks by understanding natural language instructions. For example, asking an LLM to "Translate the sentence 'Hello world' into French" is a zeroshot task. This method is especially valuable in cases where annotated training data is unavailable as it eliminates the need for task-specific fine-tuning. Zeroshot prompting shows the versatility of LLMs in handling diverse tasks by relying solely on their massive training corpora and inherent language understanding [14].

Few Shot:
Fewshot prompting is a method where the language model is given a small number of examples within the prompt to generate the desired task before asking the model to perform it. By including one or more input-output pairs as part of the prompt, the model gains contextual understanding of how to approach the task. For example, a prompt might include: "Translate the following sentences into French. Example: 'Hello' -¿ 'Bonjour'. 'Good morning' -¿ 'Bon matin'. Now translate: 'How are you?'" Fewshot prompting reduces the need for task-specific fine-tuning by enabling the model to infer task instructions from the provided examples. It is especially effective in ambiguous or complex tasks because it provides additional context for the model to generate accurate responses [16].

Chain-of-Thought:
Chain-of-thought prompting is a technique designed to improve the reasoning capabilities of LLMs by encouraging them to provide step-by-step explanations while solving complex problems. Instead of directly outputting the final answer, the prompt explicitly requests or implicitly triggers the model to break the problem into smaller, sequential steps. For example, for a simple math problem like "What is 57 plus 23, divided by 2?" a chain-of-thought prompt lead the model to first calculate 57 plus 23 equals to 80 and then 80 divided by 2 equals to 40, before arriving at the final answer. This approach enhances the models ability to handle tasks that require logical reasoning, multi-step calculations, or causal understanding. Chain-of-thought prompting significantly improve performance on tasks such as arithmetic, logical reasoning and multi-hop question answering [23].

2.2 Review of Existing Research

In this study of F., Agbavor, H. Liang [20], GPT-3 model is used to assess if acoustic features and text embeddings' can offer complementary and enhancing information for boosting the AD classification. Here acoustic features from audio data retrieved from speech and the GPT-3 based text embeddings are combined only by the help of simple concatenation. GPT-3 embeddings estimate a subject's MMSE score using three regression models like ridge regression (Ridge), support vector regressor (SVR) and random forest regressor (RFR) using RMSE score and AUC as the metric of

evaluation. It demonstrates that text embedding surpasses the traditional acoustic feature-based methods and performs on par with fine-tuned models.

The study utilizes ChatGPT-3.5 and ChatGPT-4 to identify language structures and signals pertinent to the pinpointing of Alzheimer's disease (AD). Traditional NLP methods for detecting Alzheimer's disease have several difficulties and the big problem is that there isn't enough labeled speech data from people having Alzheimer's Disease (AD). Traditional supervised learning strategy need tons of labeled data which can be hard and take a long time to gather in healthcare. To solve this problem, fewshot and zeroshot prompt learning methods let models learn with minimal labeled and unlabeled input and also enable LLMs to retrieve high-level distinguishing linguistic markers [44].

In this paper, a unique LLM-augmented literature mining approach is used to extract SDoH knowledge in AD and connect it with biological knowledge using knowledge graph creation and integration. Collected PubMed articles and used OpenAI's latest pretrained LLM GPT-4o and classic NLP methods to identify biological and SDoH entities using named entity recognition. Furthermore, they leveraged GPT's language abilities to extract SDoH and biomedical entities' potential relations and filter for co-occurrences to ensure accuracy resulting in triplets of biomedical entity, relation and SDoH entity that form an AD-related SDoH knowledge graph [48].

The research evaluates three LLM chatbots which are ChatGPT-4, ChatGPT-3.5 and Bard based on the capability to differentiate among Alzheimer's Dementia (AD) and Cognitively Normal (CN) people with the usage of spontaneous speech text later transcribed into text.Audio recordings available in the ADReSSo Challenge was chosen,well-studied dataset, the researchers employed a zeroshot learning approach where two types prompts were used: an initial query followed by a detailed chain-of-thought prompting. The chatbots' outcomes were categorized into three classes: "AD," "CN" or "Unsure." and using metrics such as accuracy,precision, sensitivity, specificity and F1 score, performance was evaluated.Bard performed best at identifying AD achieving 89% recall and F1 score of 71% but misclassified CN as AD with high confidence. GPT-4 was better at identifying CN by achieving the highest true-negative rate 56% and 62% F1 score. But three chatbots are not currently sufficient for clinical use although they can identify surpassing chance levels [36].

Recently many investigations show that large language models (LLMs) like GPT-4 not only showcase amazing competencies in Natural Language Processing (NLP) tasks but also in different professional benchmarks. In this paper, we investigate the capabilities of LLMs like GPT-4 to surpass AI tools used traditionally in dementia diagnosis by designing effective prompt template including information about a subject such as cognitive test scores and other features like biomarkers using fewshot learning to test GPT-4's ability to diagnose dementia by comparing its performance to the traditional AI tools and doctors using real clinical datasets like ADNI and PUMCH with five traditional AI models like CART, Logistic Regression, RRL, Random Forest and XGBoost on dementia diagnosis using accuracy as a metric. Results show that GPT-4 currently does not exceed conventional AI tools' effectiveness where GPT-4 shows promise in handling limited data providing readable

explanations which often align with doctors' reasoning but struggles with complex tasks, input sensitivity and lacks fine-tuning capabilities making it less effective than traditional AI models [30].

This study discusses the exploitation of large language models (LLMs) specially chatGPT GPT-3.5 model and natural language processing (NLP) technologies in a real-time chatbot application combined with machine learning to provide interpretable predictions of cognitive decline based on using spontaneous speech. The pipeline includes (i) Nlp-based prompt engineering along with reinforcement learning from human feedback (RLHF) for extraction of data; (ii) stream-based data processing techniques which includes linguistic-conceptual features for the analysis of natural language and feature engineering and selection methods like correlation and variance thresholding to enhance performance with optimal outcomes achieved through the ARFC algorithm; (iii)classification in real-time and (iv) the dashboard of explainability which achieves over 80% accuracy with recall of 85% for mental decline offering a flexible and personalized diagnostic tool [32].

This paper explores strong potential of recent large language models (LLMs) to annotate domain-specific data related to dementia in a fast cost effective way which is traditionally done by humans which is very time-consuming and expensive. The study compares LLM-generated annotations with human annotations on words/phrases in ambiguous and domain-specific texts taken from "Healing Well" forum of family carers of people with dementia. Both LLM and human raters used an annotation scheme based on the eDEM-Connect ontology to label named entities (NEs) and relationships in dementia texts.This study compared manual annotations by two raters with LLM-based annotations generated by GPT-4o both followed the IOB tagging format while using a Chain-of-Thought (CoT) prompting technique with examples. While the LLM found twice as many instances as human raters, there was moderate overlapping between them an the agreement between the LLM and the expert was moderate with 0.48 denoting kappa score and 0.87 F1-score but LLM often made mistakes for example misclassifying possessive pronouns as "Person with Dementia" (PwD) or "Family-Carer" (FC). It also struggled with complex and context-dependent information unlike expert raters who were more accurate due to their experience.The findings suggest that LLMs can improve the task of annotating data with the help of human expertise to refine the results [50].

M., Ribeiro et al. [46] discusses an explainable LLM method that identifies AD-related lexical components and prioritizes them for LLM decision-making which is SLIME (Statistical and Linguistic Insights for Model Explanation) and it focuses on word-level attributions. A Cookie Theft picture description task transcription dataset in English was used to construct this approach. Bidirectional Encoder Representations from Transformers known as BERT identified text-based representations as either Alzheimer's disease (AD) or control groups. As well as, employed Word Count (LIWC), Integrated Gradients (IG), Linguistic Inquiry, statistical analysis to select significant lexical features and decide the most important ones to influence the model's decision. It achieved an accuracy rate high as 87% in an experiment of 5-fold cross-validation.

This paper detects dementia via textual input with in-text pause encoding. Three distinct pauses (short, medium and long) are taken from the audio and encoded with the matching text using various encoding techniques. This strategy is innovative and the in-text pause encoding approaches have yet to be evaluated on the Pitt Corpus Cookie Theft dataset. BERT-base-uncased model is used as the base for feature extraction. The findings demonstrated that the Pitt Corpus Cookie Theft method was successful in enhancing performance relative to the baseline. Using only in-text encoding improves accuracy and f1-score to 0.74 and 0.77 respectively. For other encodings, the disparity is significantly greater. The model performed best overall with accuracy and f1-score of 0.86 and 0.88. Compared to baseline, the f1-score increases by almost 2 [49].

This research explores new ways to detect dementia by analyzing how people use language in their speech by using advanced machine learning and deep learning techniques to study language patterns. Mainly based on the "Cookie Theft" picture description test and speech transcripts collected from DementiaBank specifically the Pitt Corpus and ADReSS challenge datasets, it executes fine-tuning machine learning models. It compares traditional machine models like Random Forrest, Logistic Regression and Support Vector Machines with more recent advanced models including BERT, DistilBERT, RoBERTa, Mistral and Llama and performance was evaluated based on accuracy, AUC and F1 scores. Random Forest excelled among classical ML models while BERT showed excellence with advanced fine-tuning methods.The study shows Large Language Models (LLMs) exhibits great possibilities in identifying dementia with high accuracy if they are improved with techniques like Low-Rank Adaptation(LoRA) and Quantization(QLoRA) [38].

U., Lokala, et al. [41] explore the challenge of diagnosing mild cognitive impairment (MCI) among senior citizens which is the early sign of dementia using artificial intelligence combining machine learning methods particularly knowledge-guided and large language models (LLMs) integrating linguistic and demographic features from description tasks based on two commonly used pictures like cat rescue and cookie theft. Traditional machine learning and deep learning models were used in the training testing process including Logistic Regression (BOW + LR), SVM, CNN, RoBERTa, BiLSTM with Attention, BERT, BiLSTM,and BioBERT. Aiming to differentiate between three stages of MCI for example healthy, MCI and possible MCI, Cognitive Cross-Domain Attention (CCDA) was developed for a GPT-3.5 model helping to focus on key information like signs of cognitive decline by using external knowledge from sources like ConceptNet which makes the model better at identifying important words and concepts linked to cognitive decline. This approach showed effectiveness in identifying MCI levels on a vast dataset achieving a significant AUC of 0.81% and 0.73% F1 score.

This study introduces a multimodal deep learning model combining speech, language and vision data to estimate MCI(mild cognitive impairment) and it uses a mid-level fusion approach with co-attention mechanisms to combine three different types modalities at the embedding level for the identification of potential interactions between different data within cross-transformer layers effectively capturing relationships across modalities. Tested on the I-CONNECT dataset containing semi-

structured conversations recorded via webcam between 75+ years old and interviewers where BERT model generates contextual word embeddings based on transcribed speech, Wav2Vec base model converts original audio waveforms into speech embeddings, the proposed method achieved an average AUC of 85.3% outperforming unimodal methods which showed 60.9% while bimodal approach achieved 76.3% of AUC score. The study demonstrates that integrating complementary information from multiple modalities enhances prediction accuracy [45].

T., Mo, et al. [35], the authors focus on improving the efficiency of diagnosis of Alzheimer's disease (AD) through speech analysis by proposing a new framework that uses large language models (LLMs) to enhance reasoning and systematically identify linguistic deficits and they are applied to improve AD detection by feeding them into a BERT-based models (like Albert and ClinicalBERT),BERT+RA3 (a baseline using only a few linguistic attributes) and BERT+RA13 (the new model with 13 linguistic attributes for reasoning). Albert+RA13 which outperformed previous models in accuracy, precision and recall showing an 8.5% improvement in accuracy and F1 score using speech transcripts from a dataset (ADReSS) compared to traditional methods by used 13 linguistic attributes . The results highlight that incorporating detailed linguistic profiles helps distinguish between healthy individuals and Alzheimer's patients more effectively.

This work presents a Language for Pre-Screening and Cognition Assessment Model (PST-LCAM) designed for the early identification of cognitive deterioration, specifically dementia through a hybrid methodology that integrates Natural Language Processing (NLP) with Role and Reference Grammar (RRG). It utilizes linguistic besides psycholinguistic analysis of utterances from the DementiaBank dataset to detect early signs of cognitive deficits. The methodology correlates participant outputs with the Global Deterioration Scale (GDS) to deliver qualitative and quantitative pre-screening results [33].

This study seeks to analyze the capabilities of large language models (LLMs) in detecting cognitive decline such as mild cognitive impairment (MCI) from real electronic health record (EHR) notes from patients aged 50 and above by comparing the effectiveness of LLMs (like GPT-4 and Llama 2 on HIPAA-compliant cloud-computing platforms) with traditional machine learning models . The study found that GPT-4 performed better than Llama 2 but none of the LLM outperformed traditional models but an ensemble model consisting of Llama 2 ,GPT-4 and traditional models including a neural network with a hierarchical structure incorporated with attention mechanisms and XGBoost excelled most achieving high precision (90.3%), recall score of (94.2%) and F1-score (92.2%). Error analysis revealed that some mistakes were similar among the models but the majority of errors were unique to each model which denotes their distinct error patterns.In conclusion, using a combination of LLMs and traditional models improved the diagnostic performance with higher accuracy [37].

According to J., Wang et al. [51], the research employs Logistic Regression, Multi-Layer Perceptron (MLP) and XGBoost as supervised learning models to evaluate electronic health records (EHRs) for predicting the risk of Alzheimer's Disease and

Related Dementias (ADRD). The high confidence subgroup which is created by averaging confidence levels across these SL models significantly enhances ICL performance. The 7B and 70B versions of the LLaMA2 models are used as the large language models (LLMs) in the pipeline. The 70B model demonstrates skill in low confidence samples when SL models fail to perform in the inference process. The 7B model is further examined in the study and optimized using dedicated medical datasets such as Asclepius-7B.Such model optimization corresponds to the ADRD task resulting in better performance. The results here show the synergistic effectiveness of both LLMs and SL models in enhancing the accuracy of ADRD risk prediction [51].

The research investigates how various machine learning techniques were exploited to recognize and label cognitive decline through the examination of acoustics features derived out of recorded speech. It uses three classifiers, Logistic Regression (LR), CatBoost (CAT) and Support Vector Machines (SVM) and it uses stratified layered 10-fold cross-validation to perform a full evaluation of the model. Shapley Additive Explanations (SHAP) is used to select the features and the process is improved several times to determine the most significant features. It examined 266 participants (Italian and Spanish speakers) and identified significant disparities in MMSE scores, age and educational attainment among the groups [31].

According to the research, Delta-KNN quantifies the relative benefits of training examples using a delta score, which estimates model performance improvement. It dynamically picks appropriate input representatives with a KNN-based retriever. Existing In-Context Learning (ICL) approaches for AD detection are limited, but this method achieves state-of-the-art (SOTA) performance. It works across LLMs and is model- and prompt-agnostic. Using OpenAI embeddings as the text encoder for enhanced retrieval performance, the Delta Matrix is built by driving the LLM in zeroshot and one-shot circumstances. It also used text-understanding-based ConE Selection to reduce Conditional Entropy (ConE) in choosing instances. These developments help Delta-KNN to shine in AD detection and other text-based applications [54].

Based on the study, although MMSE and CDR are resource-intensive, they may overlook early indications of Alzheimer's disease (AD), which is why early detection is so important. Since symptoms of Alzheimer's disease include slurring, pauses and repetition in speaking, speech-based approaches show potential. Despite achieving approximately 74% accuracy, acoustic features (like eGeMAPS) and deep models (like Wav2Vec2 and HuBERT) frequently do not possess task-specific reasoning. Despite using language models such as BERT and ChatGPT on speech transcripts, their performance remains low without fine-tuning. By directing intermediate steps, Chain-of-Thought (CoT) prompting improves LLM reasoning, making it more accurate and easier to understand. It utilises Whisper for transcription and CoT-augmented LLaMA3.2-1B with LoRA fine-tuning for the detection of Alzheimer's disease from image descriptions. Their method surpasses previous methods and demonstrates the potential of reasoning-enhanced LLMs for scalable, non-invasive AD screening, achieving an 87.5% F1-score [58].

Lin and Washington [40] proposed a multimodal deep learning framework for dementia classification using audio recordings and text transcripts from the Pitt Cookie Theft dataset. Their approach employed Wav2Vec for acoustic embeddings and Word2Vec with LSTM networks for linguistic modeling. The study evaluated four dataset configurations, including original and augmented versions. Experimental results showed that models trained without data augmentation achieved approximately 60% accuracy and around 70% AUROC, whereas the introduction of synonym based text augmentation improved performance substantially, reaching nearly 80% accuracy and up to 90% AUROC. These results demonstrate the presence of the intensive impact of augmentation on model performance. Against this, our work does not involve data augmentation and thus focuses on assessing dementia detection in the conditions of naturally limited data, so that measured performance is a result of inherent model ability and not artificially inflated training sets.

The methodology[55] employs pretrained large models without training them from scratch. GPT, BERT and CLIP are used to extract textual representations, whereas CLAP is used to extract acoustic representations. In the case of multimodal modeling, audio and text embeddings are initially obtained separately and subsequently combined with each other using element wise addition and a cross attention based classification model is used. A context based setup is used in which balanced control and dementia samples are provided as contextual examples during inference. Experiments are conducted under patient level stratified splitting, ensuring no speaker overlap between training and testing data.On the Pitt Corpus, the best performance is achieved using GPT text embeddings combined with CLAP audio embeddings on raw (non expert annotated) transcripts. This configuration achieves an F1-score of 83.33%, outperforming text only and audio only baselines. Notably, raw transcripts consistently outperform expert-annotated CHAT data, indicating that large pre trained models can extract meaningful linguistic patterns without costly manual annotations.

This paper evaluates dementia detection with speech collected from the Cookie Theft task. Acoustic features are extracted using wav2vec 2.0, leveraging all 24 contextual transformer layers. Instead of selecting a single layer, the proposed Two-Step Attention-Based Feature Combination (TSAC) framework learns channel-level and layer-level attention weights to combine these layers into a single acoustic representation.For linguistic modeling, BERT embeddings are extracted from transcripts generated by automatic speech recognition. Acoustic and linguistic features are then fused using a cross-attention mechanism (TSAC-ATT), allowing each modality to attend to the other before classification. All experiments use speaker-independent cross-validation, ensuring strict patient-level separation.On the Pitt Corpus, the acoustic-only wav2vec 2.0 system achieves the strongest performance, reaching an F1-score of 82.54%, outperforming linguistic-only and multimodal fusion configurations. The TSAC-ATT fusion model also performs strongly but does not surpass the acoustic-only setup on this dataset. These results demonstrate that high-quality self-supervised acoustic representations can capture dementia-related markers effectively within the Pitt Corpus.[57].

Ortiz et al.[27] proposed a deep learning multimodal architecture that integrates

textual and audio information extracted from Pitt Corpus recordings. Their experiments demonstrated that textual models significantly outperformed acoustic and multimodal variants. Specifically, the text-only model using BERT embeddings combined with a bidirectional LSTM achieved an accuracy of 90.36%, compared to 73.49% for the audio-only CNN model and 86.65% for the multimodal fusion approach. These results highlight that within the Pitt Corpus, linguistic information alone is sufficient to capture salient cognitive markers of dementia often outperforming more complex multimodal pipelines.

Shakeri et al.[47] investigated whether classical machine learning models could achieve competitive results on the Pitt Corpus without relying on deep or multimodal architectures. Using TF-IDF–based linguistic features extracted from Pitt Corpus transcripts, the authors evaluated multiple classifiers with cross-validation. Their results showed that Support Vector Machine (SVM) achieved the highest accuracy on the Pitt Corpus at 90%, closely matching state-of-the-art deep learning approaches. Random Forest followed with 88% accuracy, while Logistic Regression and Naïve Bayes achieved 86% and 84%, respectively. Although KNN lagged behind at 74%, overall AUC scores remained high (up to 0.94), indicating strong discriminative ability across models. These findings demonstrate that carefully designed classical NLP pipelines can rival deep learning models on the Pitt Corpus.

2.3 Summary of Key Findings

This section is organized into three main parts: Key Results which highlights the most significant outcomes of the research, Patterns and Trends identifies themes and insights observed throughout the study and Scope for Improvements discusses potential areas for further enhancement and future work.

2.3.1 Key Results:

- GPT-3 embeddings combined with acoustic features improved dementia classification using Ridge Regression (Ridge), Support Vector Regression (SVR) and Random Forest Regression (RFR), demonstrating the effectiveness of multimodal data.
- ChatGPT-3.5 and GPT-4 identified linguistic patterns relevant to AD using prompt based learning, reducing the need for large labeled datasets and improving early diagnosis capabilities.
- GPT-4o was able to successfully extract Social Determinants of Health (SDoH) knowledge linked to Alzheimer's disease from PubMed and integrate it into knowledge graphs. In addition to this, it achieved a moderate level of agreement with human annotators in dementia-specific text labeling (kappa = 0.48, F1-score = 0.87).
- In the comparison of traditional artificial intelligence and LLM chatbots, Bard had the highest recall (89%) for Alzheimer's disease (AD) recognition, whereas GPT-4 was more accurate in identifying between persons who were cognitively normal. However, when applied to clinical datasets (ADNI and PUMCH)

GPT-4 could not achieve diagnostic accuracy that was superior to that of common AI models such as XGBoost and Random Forest algorithms. Through its utilization, interpretability and reasoning were improved.

- An accuracy rate of over 80 percent was attained using chatbots powered by artificial intelligence that utilized GPT-3.5, Reinforcement Learning with Human Feedback (RLHF) and linguistic feature selection. This demonstrates the promise of speech based Alzheimer's disease diagnosis. In-text pause encoding with BERT considerably enhanced classification accuracy (0.86) and F1-score (0.88).

- The combination of Statistical and Linguistic Insights for Model Explanation (SLIME) and BERT resulted in an increase in the accuracy of AD categorization in the range of 87%. In addition, the integration of text, voice and vision data through the use of BERT and Wav2Vec improved the categorization of Alzheimer's disease (AUC = 85.3%), bettering the performance of unimodal and bimodal models.

- LLaMA2-7B and 70B improved the risk of predicting Alzheimer's dementia disease when paired with Logistic Regression, XGBoost and MLP. Additionally, the performance of the model was improved by specific datasets such as Asclepius-7B.

- Prior multimodal deep learning approaches on the Pitt Cookie Theft dataset demonstrate high sensitivity to data augmentation. While augmentation strategies (e.g., synonym replacement) can boost performance up to 80% accuracy and 90% AUROC, baseline models without augmentation often perform substantially lower (60% accuracy), indicating potential performance inflation due to synthetic data.

- Results from pretrained large-model–based approaches show that feature extraction without fine-tuning can achieve strong performance. Combining GPT-based textual embeddings with CLAP acoustic embeddings yields an F1-score of 83.33%, outperforming unimodal baselines and demonstrating the effectiveness of representation-level multimodal fusion.

2.3.2 Patterns and Trends:

- For the purpose of identifying linguistic and cognitive signs of Alzheimer's dementia (AD), Large Language Models (LLMs) are increasingly being deployed. These LLMs frequently outperform standard rule based natural language processing (NLP) methods. These models make it possible to perform automated screening for dementia by evaluating data gathered from spontaneous speech and text, thereby lowering reliance on structured diagnostic methods.

- The current study environment recommends that there should be a move toward merging many sorts of data, such as text, audio and pictures to enhance the correctness of dementia disease categorization. To develop a deeper insight into cognitive decline, multimodal techniques are utilized. These approaches involve the combination of LLMs such as BERT and GPT-4 with speech-processing models such as Wav2Vec and CNN based audio classifiers.

- The objective of making AI based dementia diagnosis more understandable and interpretable is attracting an increasing amount of interest. Important cognitive and language signals are prioritized through the use of techniques such as Statistical and Language Insights for Model Explanation (SLIME) and Cognitive Cross-Domain Attention (CCDA). Taking this step will allow us to guarantee that the predictions made by AI are in line with clinical rationale.These methods make models more reliable and help gain trust in the medical field

- LLMs are limited in their efficacy in the healthcare industry due to the fact that they encounter challenges like reliability, biased data and the requirement for expert reviews. Despite the fact that LLMs are able to demonstrate their success, this is still the case. The presence of human oversight is essential due to the fact that the model may make errors, there may be concerns over privacy and individuals have varying ways of communicating. It is for this reason that LLMs are more useful as supplementary tools than as definitive answers to diagnostic questions.

2.3.3 Scope for Improvements

- LLMs have significant potential for dementia diagnosis particularly in widening model comparisons beyond BERT or GPT-3 topologies. LLaMA and PaLM are recent models that can be examined to determine their efficacy in cognitive decline studies. In addition, the incorporation of multimodal data including speech patterns, facial expressions and physiological signs can provide a more comprehensive approach to the identification of symptoms associated with dementia.

- Because of the limited number and variety of the datasets that are currently available, it is an absolute necessity to enhance the generalizability of models. An increase in the amount of training data that incorporates a wider variety of linguistic, demographic and cognitive characteristics could potentially improve the relevance of models for a variety of different populations. It is of the utmost importance to study methods that safeguard privacy in order to secure the confidentiality of sensitive health data such as speech recordings without affecting the accuracy of the data. This can be accomplished within the context of protecting the privacy of voice recordings.

- Due to the fact that errors in transcriptions could potentially lead to incorrect interpretations of cognitive impairment, improving transcription precision and contextual comprehension in speech-to-text models is a critical element. Enhancing resource efficiency is essential because of the computational intensity of LLMs. Enhancing their efficiency and accessibility would augment their usability in healthcare environments.

- It is important to improve the evaluation systems because the existing assessment tools fail to reflect the complexity of the AI based dementia diagnosis. Increased standardization of examinations would be facilitated by development of more comprehensive benchmarks and would result in increased clinical trust. Moreover, the use of more sophisticated learning methods such as zeroshot or

fewshot learning can help reduce reliance on large labeled datasets and therefore enhance the flexibility of models to new conditions with only a small quantity of data.

- Guaranteeing that data collection processes are of quality and homogeneous is necessary to enhance model accuracy. The application of standardised criteria to record and mark linguistic and cognitive signs would enhance the accuracy and repeatability of dementia identifying models to enable more productive AI controlled cognitive assessments.

Chapter 3

Requirements, Impacts and Constraints

This section explores effects of the research covering societal impact to understand its influence on communities, environmental impact to assess ecological consequences and ethical issues to address moral concerns. It also includes a Project Management Plan detailing how the project was organized and executed and Risk Management which identifies potential challenges and solutions.

3.1 Societal Impact

Impact on Society:
This technology can analyze speech to detect signs of memory problems and support early intervention. People at risk can enjoy a better quality of life, while their families and caregivers face less stress and work. The study aims to enable early detection especially in areas with limited medical resources by using automated dementia tests.

Impact on Health:
By offering an AI tool to help with dementia detection, it helps lighten the work of medical professionals. This method reduces human mistakes, speeds up healthcare decision-making and makes it easier to identify cognitive problems.

3.2 Environmental Impact

Sustainability:
By employing digital approaches for the purpose of data collecting and analysis, our study contributes to the maintenance of sustainability. Conventional clinical examinations frequently rely on documentation that is written down on paper and in-person evaluations both of which need the availability of physical resources and transportation. By digitizing the whole process,we reduce paper consumption resulting in less deforestation and less waste generation and minimize travel-related carbon emissions.

Computational Energy Consumption:
To be able to execute large language models (LLMs) and vision language models, there must be a significant amount of processing capacity available such as high-performance computing clusters and cloud-based infrastructure. These activities cause a lot more carbon dioxide to be emitted as well as energy consumption. This is why we would like to use eco-friendly, low-impact GPUs and cloud storage instead of more heavily damaging ones for the earth. Resources with reduced carbon emissions and better energy efficiency will be given more importance if we are to reach this target. This will thus mean that our study will remain environmentally conscious without sacrificing its capacity to run high-performance computation.

3.3 Ethical Issues

- Consent and Confidentiality: Ensure participants provide informed consent are aware of data usage and that personal data is anonymized and securely stored to protect privacy.
- Data Bias and Fairness: Ensure the datasets are diverse and balanced so the predictive models are fair and accurate.
- Ethical Use and Impact on Participants: Prevent misuse of the models in healthcare and protect participants from any psychological harm during the study.

3.4 Economic analysis

Dementia often goes unnoticed until it becomes severe resulting in high medical costs and difficult decisions for patients and their families. A mobile app or web-based platform for primary dementia screening can help detect early signs more easily than traditional hospital-based methods. Early screening allows patients to make informed health and lifestyle choices potentially lowering future treatment costs. If brought to the medical market, the platform could also generate revenue through clinics and healthcare providers while improving access to early dementia care.

To keep the platform sustainable and improving over time, it needs to generate revenue. Ongoing resources are required to maintain secure systems, update models, protect data privacy and comply with medical and technical standards. Sustainable earnings enable regular system maintenance, model refinement and scalability while preserving reliability and accessibility.

Revenue can be generated through institutional adoption, where clinics, hospitals and healthcare providers use the platform as a preliminary screening or decision-support tool. Additional models may include licensed access for healthcare organizations, integration with clinical management systems or subscription based professional use. Such approaches support sustainable operation while allowing patients and caregivers to benefit from a streamlined and accessible early screening solution.

3.5 Project Management Plan

The chart below presents the revised timeline of this thesis, which is structured into three phases: Pre-Thesis 1, Pre-Thesis 2 and the Final Thesis. The period from June to September was deferred, after which the thesis work was resumed.

Pre-Thesis 1 (Oct–Jan) involves preliminary research activities, including topic selection, problem formulation and an extensive literature review. During this phase, the research objectives are defined and the P1 Report is prepared and submitted.

Pre-Thesis 2 (Feb–May) focuses on dataset collection, methodology and design selection, model development planning and experimental setup. This phase concludes with the preparation and submission of the P2 Report, along with the poster presentation.

The Final Thesis phase (Oct–Jan) includes implementation of the proposed framework, experimental evaluation, performance analysis, final system refinement and completion of the final thesis report. The process concludes with the final submission and presentation.

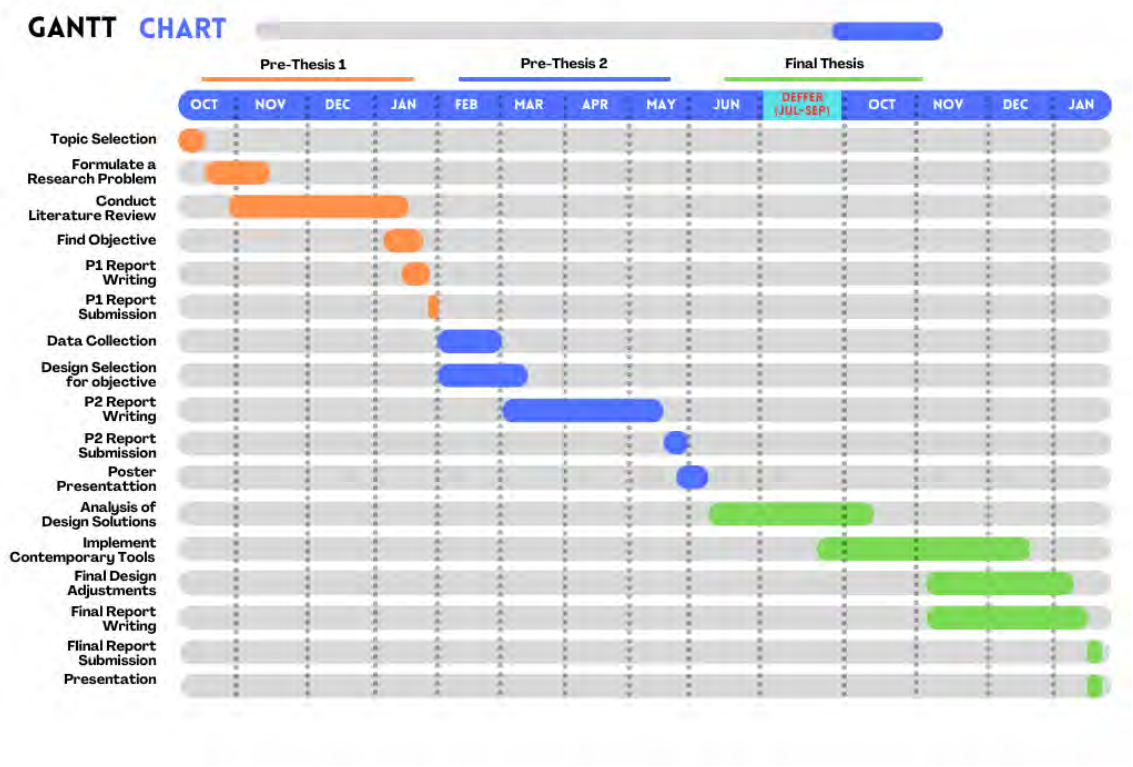


Figure 3.1: Thesis Planning

3.6 Risk Management

A key risk in this study is the inability to implement all the three objectives due to data limitations, integration challenges or unforeseen issues during model development. Then the study will focus on optimizing the two primary modalities to ensure the best possible predictive accuracy with the existing resources. Any challenges that are associated with the incorporation of new modalities will not be able to hamper development if linguistic and auditory analysis are improved. By utilizing modern natural language processing and audio approaches, the study will instead refine existing procedures in order to reduce the likelihood of delays. Even if the third modality cannot be integrated because there is insufficient data, the research will still produce significant insights by concentrating on the two modalities that are accessible. This will ensure that the research makes contributions that are relevant to the identification of dementia.

Chapter 4

Proposed Methodology

The following part provides an overview of methodology explaining how the objectives are going to be achieved, different models, data collection process with detailed cleaning and transformation descriptions and lastly implementation of selected design.

4.1 Design Process or Methodology Overview

This paper explores the application of the multimodal methodological approach in detecting dementia by using a multimodal approach to both speech and language based on Pitt corpus of dementiabank. The general layout is to focus on the strong preprocessing, complementary feature extraction and systematic assessment that is based on both traditional machine learning models and the contemporary foundation models that guarantee credibility and clarification across modalities.

The initial approach is preprocessing of audio and transcript material. The speech recordings are processed to normalization and denoising, segmented with transcript aligned timestamps and deprived of noninformative silence to get high quality participant only speech signals. Transcripts are cleaned and matched to audio and preprocessing strategies are optimized to particular modelling goals. To study linguistic data, linguistic analysis resorts to such tools as NLTK to study lexical richness, coherence, speech rate, repetition,etc. Acoustic analysis combines handcrafted features and deep representations to capture both low level and high level speech characteristics associated with cognitive decline. Handcrafted features are extracted using Librosa, focusing on prosodic, temporal and spectral attributes such as pitch variation, pause behavior and Mel-frequency cepstral coefficients.

For predictive modeling, the study first uses the CLAP framework to get embeddings from both audio and text. These embeddings are then combined with acoustic and linguistic features to see how the features improve the representations. At the same time, traditional machine learning methods like logistic regression, random forests, support vector machines gradient boosting, etc. are used to perform supervised classification and check how well the features work.

Beyond traditional models, we also evaluate Large Language Models (LLMs) and Vision–Language Models (VLMs) in an inference-only setting, without any fine tuning.

LLMs and VLMs are tested using zeroshot, fewshot and chain-of-thought prompting as well as Guided prompts that provide clear rubrics to steer their reasoning. For VLMs, visual representations extracted from speech data are used and in the combined VLM+LLM setup, the VLM first provides visual cues which the LLM then looks at step by step, combining them with language information to predict dementia labels accurately.

Model evaluation follows a rigorous experimental protocol with patient-level data splits to prevent information leakage. Stratified train–test partitions and cross-validation are employed to preserve class balance and ensure generalization. Performance is evaluated using standard classification metrics.

In general, the design process provides a single and generalizable methodology of detecting dementia that spans handcrafted capabilities, rich multimodal representations and foundation model reasoning. Through systematic comparison of multimodal and prompt based methods in situations of consistent evaluation, the study is an exhaustive evaluation of the role of language, speech and multimodal reasoning in analysis of cognitive health.

4.1.1 Contrastive Learning Based Methodology

This research involves the use of multimodal approach to classifying dementia that involves the use of both audio recordings and textual transcripts. The Contrastive Language Audio Pretraining (CLAP) model is a model that learns audio text semantic correspondence and produces aligned embeddings in a common representation space. Acoustic and linguistic crafted features are also added to supplement the CLAP embeddings.

To reduce the dimensionality of high-dimensional embeddings, Principal Component Analysis (PCA) is applied independently to CLAP audio and text embeddings, retaining either a fixed number of components or a target variance threshold. Feature selection is applied exclusively to handcrafted features using SelectKBest, Recursive Feature Elimination (RFE) and SelectFromModel.

The dataset is split at the patient level to ensure true generalization with cross-validation (CV) and held-out test sets stratified by diagnosis. A five-fold stratified CV is performed on the CV set to evaluate model performance. Multiple classifiers, including Logistic Regression, Random Forest, Extra Trees, XGBoost, LightGBM, Linear SVM, RBF SVM, K-Nearest Neighbors, Linear Discriminant Analysis and Naive Bayes are employed to ensure robustness. This framework provides a systematic, multi-modal, patient-level approach for evaluating dementia classification.

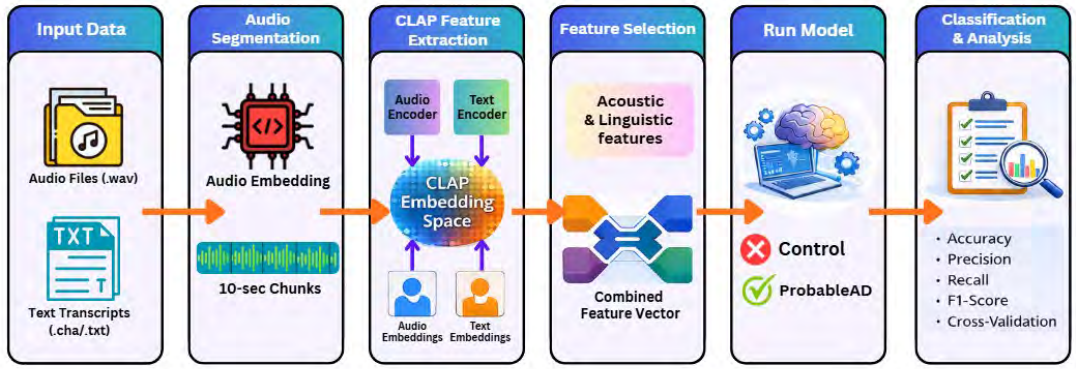


Figure 4.1: CLAP Based Pipeline

4.1.2 Large Language Model Based Methodology

This study employs Large Language Models (LLMs) to perform dementia label detection from textual transcripts using an inference-based evaluation framework. Rather than training or fine tuning model parameters, the focus is placed on analyzing how pre trained LLMs reason and make classification decisions under different prompting strategies. This design enables a systematic investigation of the influence of prompt formulation on model behavior within a clinically relevant language understanding task.

Three open source LLMs such as Qwen3-4B, GPT-OSS and Mistral are chosen because of their architectural diversity and ability to be used in instruction following and natural language inference. Models are all operated in a frozen state such that observed variations in performance are due to variations in prompting strategy and not parameter adaptation or learning dynamics.

The dementia detection task is defined as binary classification problem and the transcripts are assigned individually one of the following conditional groups ProbableAD and Control. There is no inter sample contextual information exchange, which supports the study of the different levels of contextual guidance and reasoning structure. There are five prompting strategies that are used, Zeroshot, Fewshot, Chain-of-Thought (CoT), Guided and Guided Prompt with Fewshot demonstrations. By the Zeroshot prompting, the model is provided with task level instructions only and its own semantic knowledge can be evaluated. In fewshot prompting The labels are added to the examples to permit in context learning whereas Chain-of-Thought prompting promotes explicit reasoning before the model classifies the linguistic pattern correlated with cognitive deterioration. This strategy is further expanded by the Guided Prompt with Fewshot strategy that includes structured clinical prompts as well as example demonstrations in order to restrict decision making and minimize output variability.

Model predictions are compared against ground truth dementia annotations using standard classification metrics, including Accuracy, Precision, Recall and F1-score. Performance analysis is conducted across two complementary dimensions, prompt

based evaluation, which assesses the relative effectiveness of different prompting strategies and model based evaluation which compares Qwen3-4B, GPT-OSS and Mistral under identical prompting conditions. This dual analytical framework enables identification of interactions between model architecture and prompt design.

The study offers a clear understanding of the effects of the various prompting strategies on reasoning behavior and classification performance of Large Language Models on clinical language understanding tasks.

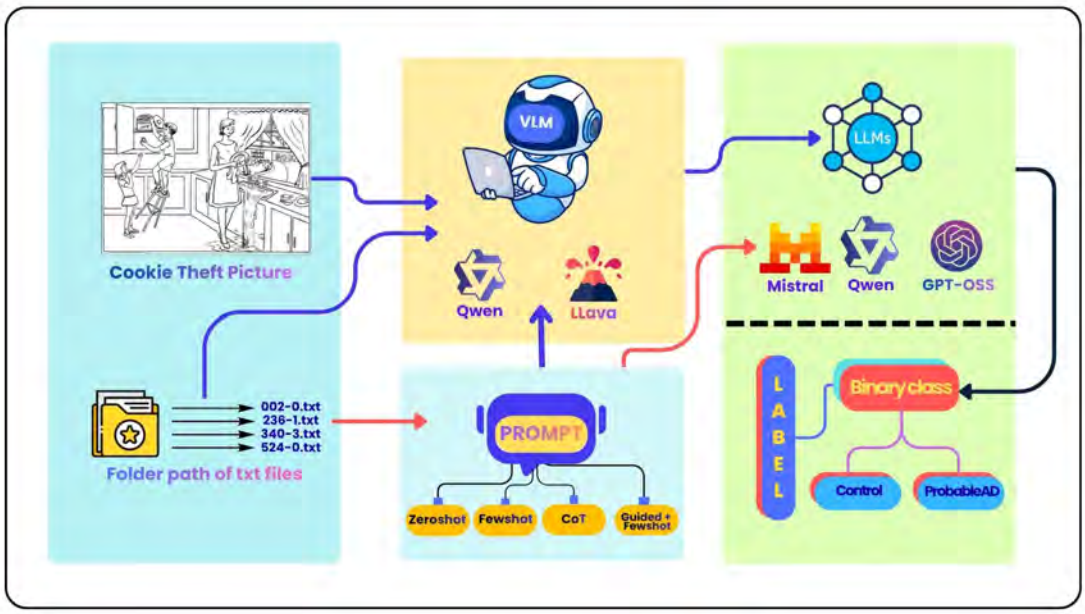


Figure 4.2: Workflow of LLM and VLM

4.1.3 Vision Language Model Based Methodology

The paper explores the application of two multimodal reasoning approaches: only a VisionLanguage Model (VLM) and a hybrid LLM + VLM approach. The two methods are to examine cognitive impairment through a combination capitalizing on the visual evidence based on speech data and related textual information. The experiments are done within an inference only framework without fine tuning any model to study the effect of multimodal reasoning and prompt design classify performance of influence.

Vision Language Model-Only Approach

In the VLM only model, dementia classification is realized in direct mode by applying to the description of cookies theft of visual and textual information by jointly using Vision Language Models. Two open source VLMs, LLaVA-v1.6-vicuna-7B-hf and Qwen2.5-VL-3B-Instruct are used because of their well developed multimodal reasoning. A sample is a visual representation of a speech data of a participant along with the textual context of the speech of the participant. These inputs are also concurrently made to VLM that facilitates cross modal attention of the visual

cues and language patterns that are suggestive of cognitive impairment.

The task is formulated as a binary classification problem, where each instance is independently labeled as ProbableAD or Control. No contextual information is shared across samples reflecting realistic forensic decision-making conditions. Different prompting strategies are applied to guide multimodal reasoning, including Zeroshot, Fewshot and Chain-of-Thought (CoT) prompting. Zeroshot prompting provides only task-level instructions, Fewshot prompting incorporates labeled image–text demonstrations of some patients and CoT prompting explicitly encourages step-by-step reasoning across both modalities prior to final prediction. For each experiment, the visual input and constructed prompt are passed to the selected VLM and the generated output is parsed to extract the predicted label. Predictions are normalized into a binary format to ensure consistent evaluation.

Large Language Model + Vision Language Model Hybrid Approach

The hybrid LLM + VLM approach extends the VLM-only framework by introducing an additional language reasoning stage. In this setting, the VLM is first used to analyze the visual input and generate an intermediate textual description of the cookie theft image. This generated multimodal explanation is then passed to a Large Language Model (LLM) which performs higher-level reasoning and produces the final dementia classification. The twostage design allows the separation of the visual comprehension and language reasoning. The VLM is dedicated to the extraction of salient visual semantic cues of the information whereas the LLM focuses on the analysis of the speaking style of the patient, his/her linguistic patterns and discourse features to identify the hints of the cognitive deterioration and the lack of information that the image of the cookie theft has. This strategy should be used to improve the interpretability and decrease ambiguity by enabling the final decision to be informed by explicit reasoning as opposed to direct multimodal prediction perse.

Both approaches are compared with ground truth forensic annotations of Alzheimer by predicting the labels and comparing them with the forecasted label using conventional classification measurements such as Accuracy, Precision, Recall and F1-score. The performance analysis is carried out in order to compare the performances of direct multimodal reasoning (VLM-only) and hierarchical reasoning (LLM + VLM). All configurations have the same datasets, inference procedures and measures of evaluation to guarantee rigor in the experiments. Every preprocessing, timely construction and output parsing operation has been clearly established in the codebase.

Overall, this methodology establishes a structured comparative framework for forensic Alzheimer detection using multimodal foundation models. By contrasting VLM-only reasoning with hybrid LLM + VLM reasoning, the study provides insight into the role of hierarchical language reasoning in enhancing multimodal clinical classification performance.

4.2 Models Overview

4.2.1 Contrastive Language–Audio Pretraining (CLAP)

Contrastive Language Audio Pretraining (CLAP) is a multimodal base model that trains on large scale contrastive learning of audio and natural language joint representations. CLAP aligns semantically corresponding audio text pairs in a shared embedding space by maximizing similarity between matched pairs and minimizing similarity between mismatched ones. This training strategy enables the model to learn rich, transferable audio representations without relying on task specific labels [22].

The CLAP system comprises of an audio encoder and a text encoder which are trained together with a contrastive goal. The audio encoder is used to decode high level acoustic content in raw audio or spectrogram based inputs and the text encoder is used to encode natural language descriptions into semantic embeddings. Using extensive sets of loosely monitored audio text data, CLAP is able to extract low level acoustical features as well as high level semantic and linguistic features of the audio signals. One of the benefits of CLAP is its open vocabulary learning property that enables successful zeroshot and fewshot generalization to unseen types of audio. In comparison with the traditional audio models that are trained on closed-set labels, CLAP enjoys the benefit of natural language supervision thus being highly flexible to downstream tasks. These characteristics render CLAP especially appropriate in speech-based applications in which the semantic content, prosody and temporal patterns of speech are of great importance.

4.2.2 Lagre Language Models(LLM's)

GPT-OSS-20B
GPT-OSS-20B is an open weight large language model with approximately 20 billion parameters, designed for general purpose language understanding and generation. It employs a mixture of experts transformer architecture, which activates only a subset of parameters per token, enabling efficient inference on standard hardware. The model is trained with large scale distillation and reinforcement learning techniques to perform tasks such as instruction following, reasoning, chain-of-thought generation and code synthesis. Its open source release under the Apache 2.0 license allows researchers to freely inspect, adapt and fine-tune the model. GPT-OSS-20B achieves strong performance on multiple benchmarks, including mathematics, coding and reasoning tasks making it a practical choice for research and deployment [56].

Mistral 7B
Mistral 7B is a 7-billion-parameter open-weight large language model introduced to achieve high performance and efficiency through architectural innovations such as grouped-query attention (GQA) and sliding window attention (SWA). The model demonstrates superior performance over larger contemporaries such as LLaMA 2 13B and LLaMA 1 34B across diverse benchmarks in reasoning, mathematics and code generation, while maintaining reduced computational cost. According to researchers [24] also includes an instruction-tuned variant (Mistral 7B-Instruct) that

surpasses comparable chat-oriented models on both human and automated evaluations, illustrating that careful architectural design and instruction tuning can yield competitive small-to-medium scale LLMs under an open-source.

Qwen3-4B
Qwen3-4B is a member of the Qwen3 series of large language models that expand on prior Qwen architectures by unifying "thinking" and "non-thinking" modes within a single framework. The Qwen3 family includes dense and Mixture-of-Experts (MoE) variants spanning 0.6 billion to 235 billion parameters, designed to enhance performance, efficiency and multilingual capabilities across tasks such as code generation, mathematical reasoning and general language understanding. The Qwen3 models maintain open-source availability under an Apache 2.0 license, enabling reproducibility and research adoption [60]. Qwen3-4B, as a 4-billion-parameter instance, provides a strong balance between model capacity and computational efficiency for general NLP and reasoning task.
Qwen2.5-7B-Instruct
Qwen2.5-7B-Instruct is a 7-billion-parameter instruction-tuned large language model in the Qwen2.5 series developed by Alibaba[59]. It is intended to adhere to natural language instructions in a broad set of tasks such as reasoning, dialogue, summarization, as well as code generation. It is based on the dense transformer of Qwen2.5 and it is further refined on high quality instruction following datasets to increase its correspondence with human prompts. Its scale, instruction tuning and multilingual modeling allow high zeroshot and fewshot generalization and thus are applicable to both the research and practical use.

4.2.3 Vision Language Models(VLM's)

Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-Instruct is an instruction-tuned vision language model (VLM) that jointly processes visual inputs and natural language prompts for multimodal understanding and reasoning. It can combine a vision encoder with the Qwen2.5 large language model and is trained on the data of multimodal instructions, which allows it to perform well on specific tasks, including visual question answering, image understanding and document analysis. Its instruction following design allows effective generalization across diverse vision–language tasks without task-specific fine-tuning [53].

LLaVA-v1.6-vicuna-7B-hf
LLaVA-1.6 is an instruction-tuned vision–language assistant that extends the LLaVA framework with improved visual reasoning and alignment between vision and language representations [25]. By combining a pretrained vision encoder with a large language model and training on high-quality multimodal instruction data, LLaVA-1.6 achieves strong performance in visual reasoning, image-based dialogue and multimodal question answering.

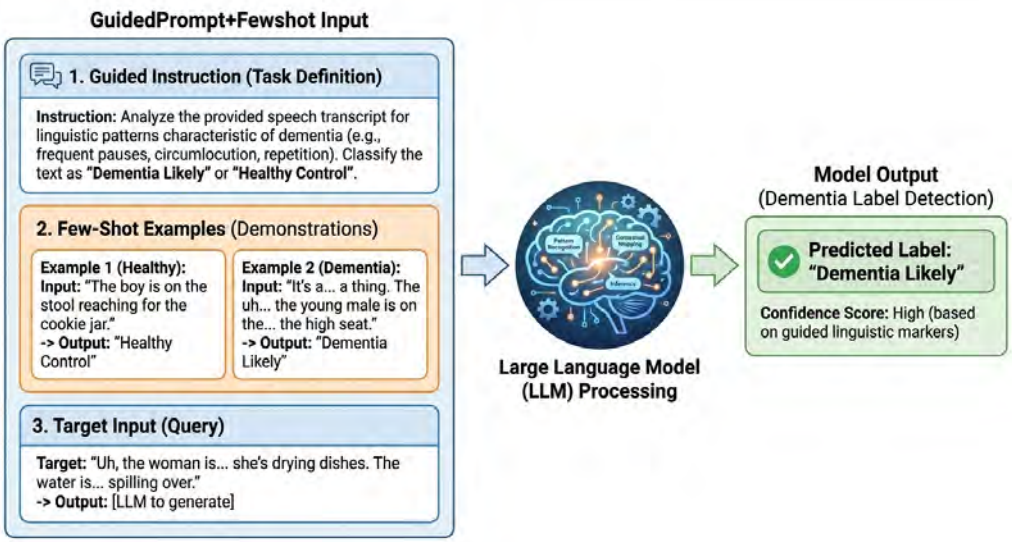


Figure 4.3: Illustration of the Guided Prompt + Fewshot prompting approach, where structured instructions and example demonstrations guide the LLM to predict the dementia label from the input transcript.

4.2.4 Prompting Techniques

In this study, we used zeroshot, fewshot and chain-of-thought (CoT) prompting strategies to assess and direct large language models. Zeroshot prompting enables the model to do jobs with no task specific examples using only pretraining knowledge and the natural language prompt. Fewshot prompting gives the model few labeled samples to show how the expected format of the task should be performed, which enhances performance on more complex or less frequent tasks. Chain-of-thought (CoT) prompting motivates the model to generate the intermediate reasoning before generating the final output and improves task performance based on multistep reasoning, logical inference or arithmetic. These prompting methods together allowed us to get a clear picture of both the raw abilities and the ability to reason of our models. [16].

Guided Prompt

Guided prompting is a structured prompting approach in which a language model is given precise and explicit instructions on how to assess the input, reason about the task and offer its final response. Instead of relying on the model to infer the task implicitly, guided prompting clearly explains the objective, emphasizes the language and cognitive aspects that must be addressed and defines a consistent output format. This structured guidance helps reduce ambiguity and limits unpredictable model behavior, leading to more stable and reproducible predictions. In the context of dementia detection, guided prompting encourages the model to focus on clinically relevant speech characteristics such as lexical diversity, sentence fluency, repetition, word-finding difficulties and overall narrative coherence. By directing attention to these indicators, the approach improves decision consistency and interpretability while avoiding changes to the model's internal parameters, making it especially suitable for sensitive healthcare-related text classification tasks.

4.3 Data Collection

The Pitt Corpus from DementiaBank is a longitudinal speech dataset consisting of the recordings from 293 participants different visits, including 1043 files with dementia and 247 healthy controls where each patient's recordings are stored in multiple files based on their visit numbers. Participants, some of whom were monitored throughout several years, completed standardized linguistic assessments including the Boston Cookie Theft visual description, word fluency, story memory and sentence building tasks. The recordings, initially in .mp3 format, were transformed into .wav files with noise reduction applied.The Dementia group consists of 309 Cookie Theft, 235 Fluency, 263 Recall and 236 Sentence recordings whereas the Control group contains 239 Cookie Theft, 2 Fluency, 1 Recall and 1 Sentence recording. All audio is transcribed using the CHAT protocol, which annotates pauses, errors and linguistic features. In the Control group, there are 243 Cookie Theft, 2 Fluency, 1 Recall and 1 Sentence files. In the Dementia group directory there are 309 Cookie Theft, 235 Fluency, 263 Recall and 236 Sentence files.

The information was gathered for the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. The study was made possible by NIH funds AG005133 and AG003705. People with probable or possible Alzheimer's disease, people without dementia and people with a variety of dementia conditions all took part.The recordings were gathered annually, allowing longitudinal tracking. Demographic details and cognitive test results (e.g., MMSE scores) are available in the Data Spreadsheet on TalkBank. Each transcript begins with a header tier

$$(e.g., @ID : eng|Pitt|PAR|73; |female|Control||Participant|29||)$$

indicating language, corpus, age, sex, diagnosis and MMSE score.

For this research, we utilized the available files from the Cookie Theft dataset.

4.3.1 Data Cleaning

The dataset contains acoustic characteristics derived from .wav files and language transcripts from .cha CHAT format files. The cleaning process addressed several difficulties. It includes textual inaccuracies, extraneous noise indicators, colloquial expressions and unnecessary symbols.

Transcript Cleaning: Only patient data from the transcripts were kept and cleaned with simple rules and patterns. The cleaning process included several steps:

For CLAP embeddings, transcripts were fully cleaned meaning no special markers were preserved allowing the model to focus solely on textual content.
In contrast, for the LLM and VLM evaluation, speaker markers such as `*PAR:` and `*INV:` were removed. Unique markers were preserved to capture paralinguistic and speech characteristics. This includes markers like `&=laugh`,long pauses `(...)`, medium pause`(..)`, short pauses `(.)`, filler words `uh` and `um`, long false starts `[//]` and short false starts `[/]`, stuttering `&+stutter`, trailing off `+...` and retrace `+<` and other CHAT annotations such as `[+ exc]` and `[+ es]`. Audio timestamps

(e.g., \x1510370_16630\x15) were removed, but all other special markers were kept unchanged for analysis.
The transcripts were carefully cleaned and then transformed into text format. The final output was structured in a CSV file.

Audio Cleaning:

To get the audio data ready for analysis, a preprocessing pipeline was used to improve signal quality and align speech with transcripts accurately. The transcript files include speaker-specific utterances with timestamp markers showing the start and end of each audio segment. For example, an entry like:

```
PAR: but the water is &+flow still flowing.  44295_46242
```

The timestamps show that the participant's spoken sentence occurs between 44,295 ms and 46,242 ms in the original audio. Using these timestamps, the relevant audio segments were extracted from the corresponding .wav files. All participant segments were then combined to create a continuous recording containing only the patient's speech, excluding the interviewer. Before segmentation, noise reduction, amplitude normalization and silence removal were applied to reduce background interference and maintain consistent signal levels. Important acoustic features—such as pitch, jitter, shimmer and timing—were preserved throughout. This preprocessing creates a clean and consistent audio dataset enabling accurate feature extraction and reliable machine learning analysis.

4.3.2 Data Transformation

The data was properly prepared for analysis and modelling by applying a variety of data transformation techniques in this study. Text embedding, category encoding and standardisation were the main approaches.

Standardization: Number features including "entryage", "onsetage"and "education" were standardised with the help of the StandardScaler in the sklearn.preprocessing module. All features were treated identically during the model training phase after this transformation. It resulted in features with a mean of zero and a standard deviation of one. The formula for determining uniformity is as follows:

$$z = \frac{x - \mu}{\sigma}$$

Where x is the original value, μ is the mean and σ is the standard deviation.

Principal Component Analysis (PCA): Principal Component Analysis (PCA) is an unsupervised linear dimensionality reduction technique that identifies orthogonal directions, called *principal components*, along which the variance of high-dimensional data is maximized. By projecting data onto a lower-dimensional subspace spanned by the leading principal components, PCA preserves as much of the original variance as possible while reducing computational complexity.
Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a dataset with n samples and d features. PCA seeks the eigenvectors $\mathbf{v}_i$ and corresponding eigenvalues λ_i of the covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad i = 1, 2, \ldots, d, \tag{4.1}$$

where

$$\mathbf{\Sigma} = \frac{1}{n-1}(\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}}) \tag{4.2}$$

is the $d \times d$ covariance matrix and $\bar{\mathbf{X}}$ is the mean vector of $\mathbf{X}$.
The eigenvectors corresponding to the largest eigenvalues capture the directions of maximum variance in the data. Ordering the eigenvectors by decreasing eigenvalues allows projection onto a lower-dimensional subspace while retaining the most significant variance. The transformed data in k dimensions is then obtained as:

$$\mathbf{X}_{\text{PCA}} = \mathbf{X}\mathbf{V}_k, \tag{4.3}$$

where $\mathbf{V}_k \in \mathbb{R}^{d \times k}$ contains the first k principal components as columns. PCA provides a compact representation of high-dimensional data, reduces noise and facilitates efficient downstream processing without requiring labeled data.

Feature Selection Methods
To improve model performance, reduce overfitting and focus on the most informative handcrafted features, three feature selection methods were applied, representing different paradigms: filter, wrapper and embedded approaches. Each method captures feature relevance in a distinct way providing complementary perspectives on feature importance.

1. **SelectKBest (Filter Method):** SelectKBest is a univariate statistical approach that evaluates each feature independently based on its correlation with the target variable. In this study, the *ANOVA F-statistic* was used to measure how well each feature discriminates between classes. Features are then ranked according to their F-scores and the top k features are retained. As a filter method, SelectKBest is computationally efficient and does not rely on a specific predictive model making it suitable for quick preliminary feature pruning. It identifies features with strong individual predictive power but does not account for feature interactions.

2. **Recursive Feature Elimination (RFE, Wrapper Method):** RFE is an iterative procedure that selects features based on their contribution to a predictive model. A base estimator (here, regularized logistic regression) is trained and features with the smallest coefficients are removed. The model is then retrained and the process is repeated until the desired number of features remains. This wrapper method considers the effect of feature combinations on model performance capturing interactions and redundancy among features. Although more computationally intensive than filter methods, RFE often provides a more refined feature subset that is optimized for the chosen model.

3. **SelectFromModel (Embedded Method):** SelectFromModel integrates feature selection directly into the training of a model that computes feature importance scores. In this study, a Random Forest classifier was used which assigns importance to each feature based on its contribution to reducing impurity across decision trees. Features exceeding a predefined importance threshold

are retained. Embedded approaches like SelectFromModel help pick relevant features while training the model, considering dependencies between features.

Missing Value Imputation
Handling missing values is an important step in data preprocessing. In this study, mean imputation was applied to all features in the CSV file, including both acoustic and linguistic data. Any missing or NaN values in a feature column were replaced with the column mean using scikit-learn's SimpleImputer(strategy='mean'). Since many machine learning algorithms cannot work with missing values, imputation ensures that all features are complete and numeric, allowing reliable model training.

Categorical Encoding: The LabelEncoder from sklearn.preprocessing was used to convert categorical labels (dx) into numerical values so they could be used in classification models. For multi-class tasks, the labels were further binarized using label binarize to make it easier to calculate metrics such as ROC AUC and precision-recall curves.

These transformations kept the important relationships and patterns in the data while making it ready for machine learning and deep learning models.

4.3.3 Data Reduction

To ensure the quality and relevance of the dataset used in the analysis, a series of data reduction steps were applied to both textual and audio data sources.

In the textual metadata, each participant was expected to have complete and consistent cognitive annotations. During preprocessing, entries lacking valid target annotations were identified and excluded from further processing. Retaining such incomplete records would have introduced noise and negatively affected model training and evaluation. However, for the classification task, no records were removed due to missing annotations, as all participants had corresponding dementia related labels. In addition, for binary classification, samples labeled as Mild Cognitive Impairment (MCI) were removed to maintain a clear distinction between cognitively healthy control and dementia classes.

For the audio modality, each transcript was paired with a corresponding .wav audio file. Several recordings were found to be unsuitable for analysis. In some cases, speech volume was extremely low or distorted, particularly for the PAR (participant) or INV (interviewer), making reliable acoustic feature extraction and speech segmentation difficult. Audio files in which the participant's voice was unclear or largely inaudible were excluded, as they negatively impacted downstream audio–text alignment and reduced the reliability of extracted speech features.These data reduction procedures resulted in a refined dataset containing only entries with complete annotations and usable audio recordings, thereby improving the overall quality, consistency and integrity of the multimodal analysis.

To create a clear binary classification problem, we removed cases labeled as vascular dementia and memory-type dementia. Labels marked as PossibleAD were replaced with ProbableAD to reduce ambiguity and Mild Cognitive Impairment (MCI) cases

were also excluded. After these steps, the dataset included 243 Control cases and 255 ProbableAD cases providing a balanced and well-defined set for dementia classification.

4.3.4 Summary of Preprocessed Data

The final datasets were prepared using a structured preprocessing pipeline to ensure data quality, consistency and alignment across modalities for dementia classification. The pipeline processed textual transcripts, audio recordings and structured metadata in a unified manner.

Textual Data:
Patient-only transcripts from the Cookie Theft task were cleaned and standardized. Depending on the modelling objective, transcripts were either fully cleaned for contrastive audio–text embeddings or retained paralinguistic markers (e.g., pauses, fillers, false starts) for LLM and VLM evaluation. All processed transcripts were stored in a structured CSV format.

Audio Data:
Audio recordings were converted to `.wav` format and preprocessed using noise reduction, amplitude normalization and silence removal. Transcript-aligned timestamps were used to extract and concatenate participant-only speech segments. Audio files with poor quality or unclear participant speech were excluded.

Feature Preparation:
Numerical demographic variables were standardized to ensure uniform scaling. Dimensionality reduction and feature selection techniques were applied to high-dimensional embeddings and handcrafted features to reduce redundancy and improve model stability. Diagnostic labels were encoded for classification.

Final Dataset Composition:
To maintain a clear binary classification setting, samples corresponding to Mild Cognitive Impairment, vascular dementia and memory-type dementia were excluded and PossibleAD labels were merged into ProbableAD. The final dataset comprised 243 Control cases and 255 ProbableAD cases, organized into text-only, audio-only and multimodal dataset variants sharing a common diagnostic target.

4.4 Implementation of Selected Design

4.4.1 CLAP and Machine Learning:

Audio and Text Processing
Audio recordings were sampled at 48 kHz and loaded in mono using **librosa**. Each recording was segmented into non-overlapping 10-second windows (480,000 samples per segment), with zero-padding applied when necessary. Text transcripts with fewer than 512 tokens were processed directly, while longer transcripts were split into overlapping chunks of approximately 256 words (25% overlap), embedded independently using the CLAP text encoder and aggregated using max pooling to

produce a single 512 dimensional text embedding.

CLAP Embeddings and Dimensionality Reduction

All experiments utilized the pre-trained **laion/clap-htsat-unfused** model, which generates 512 dimensional embeddings in a shared audio–text space. Segment-level audio embeddings were aggregated using max or mean pooling and text embeddings were pooled across chunks. To measure semantic alignment, cosine similarity was calculated between text and audio embeddings pooled together. Principal Component Analysis (PCA) was used on CLAP audio and text embeddings separately in order to reduce the dimensions. The thresholds of the explained variance (90, 95, 99) and fixed [32,64,128] were used to determine the number of components and one could choose to retain all 512 dimension. PCA-reduced embeddings were standardized using z-score normalization.

Handcrafted Feature Selection

The study uses a total of 46 features, including acoustic features capturing pitch, voice quality, formant structure and spectral characteristics, as well as linguistic features reflecting lexical richness, syntactic complexity, coherence and speech fluency. Some features from the patient's info were also taken. Feature selection was applied exclusively to handcrafted CSV features to preserve the integrity of learned CLAP embeddings. Three methods were employed:

- **SelectKBest**: A filter-based method using ANOVA F-statistics to select the top k features most correlated with the target label.
- **Recursive Feature Elimination (RFE)**: A wrapper method that recursively removes features based on their importance as determined by a base estimator (regularized logistic regression), retaining the most predictive features.
- **SelectFromModel**: An embedded method leveraging Random Forest feature importance to retain features with importance above a threshold.

Selected features were standardized using z-score normalization and used either individually or concatenated with PCA-reduced CLAP embeddings.

Patient-Level Data Splitting and Cross-Validation

To ensure patient-level generalization, the dataset was split into 80% cross-validation (CV) patients and 20% held-out test patients, stratified by diagnosis. Within the CV set, a five-fold stratified cross-validation was performed with each fold containing 64% training and 16% validation samples while preserving patient grouping. This prevented information leakage between training and validation sets and ensured unbiased performance estimates on unseen patients.

Classification Models and Hyperparameters

Ten classifiers were implemented with strong regularization to mitigate overfitting in the high-dimensional, limited-sample clinical dataset:

- **Logistic Regression (LR):**
  L2 regularization with $C = 0.01$, `lbfgs` solver, max_iter=1000.
- **Random Forest (RF):**
  50 trees, max_depth=4, min_samples_split=20,
  min_samples_leaf=10, max_features=`sqrt`.
- **XGBoost:**
  50 estimators, learning_rate=0.01, max_depth=3,
  min_child_weight=10, subsample=0.6,
  colsample_bytree=0.6, L1=1.0, L2=10.0, gamma=1.0.
- **LightGBM:**
  50 estimators, learning_rate=0.01, max_depth=3,
  num_leaves=8, min_child_samples=20,
  subsample=0.6, colsample_bytree=0.6,
  L1=1.0, L2=10.0.
- **SVM (RBF):**
  $C = 0.1$, RBF kernel, gamma=`scale`, probability enabled.
- **SVM (Linear):**
  $C = 0.01$, linear kernel, probability enabled.
- **Extra Trees (ET):**
  50 trees, max_depth=4, min_samples_split=20,
  min_samples_leaf=10.
- **K-Nearest Neighbors (KNN):**
  $n\_neighbors = 15$, distance-weighted voting.
- **Linear Discriminant Analysis (LDA):**
  `lsqr` solver with automatic shrinkage.
- **Gaussian Naive Bayes (GNB):**
  Default parameters; assumes feature independence and Gaussian likelihood.

All classifiers were trained with fixed random seeds for reproducibility. Hyperparameters were deliberately chosen for strong regularization, slow learning (for boosting) or shallow trees (for ensemble methods), to prioritize generalization over training accuracy. PCA dimensionality reduces CLAP embeddings, which is helping to avoid overfitting and maintaining informative variance. Selection of features provides handcrafted features not contribute too much that would be an excessive burden on CLAP embeddings. Patient-level splitting guarantees realistic generalization avoiding leakage from repeated participants. Strong regularization across models minimizes the training–validation gap, producing clinically valid predictions.

4.4.2 Large Language Model

Text-based reasoning was conducted using pretrained Large Language Models (LLMs) in a strictly inference only setting. The models evaluated in this component were Qwen3-4B, Mistral and GPT-OSS. All models received only textual transcripts as input, no audio signals, acoustic features or multimodal information were incorporated. Transcripts were provided as plain text following minimal cleaning performed during earlier preprocessing stages. Model behavior was governed entirely through prompt design, enabling a controlled comparison of reasoning strategies without any finetuning or parameter updates.

Multiple prompting strategies were used, including zeroshot, fewshot chain-of-thought and guided fewshot prompting with guided prompting as the main focus. Zeroshot prompting asked the model to classify transcripts using only task instructions, while fewshot prompting added a small number of labeled examples to guide the response. Chain-of-thought prompting further required the model to explain its reasoning before giving the final prediction.

Guided fewshot prompting was implemented using a structured prompt template that explicitly controls the reasoning process of the LLM while leveraging in-context examples. For each transcript, the prompt is dynamically constructed with four key elements: a task instruction defining the binary classification objective (ProbableAD vs. Control), a small set of labeled transcript examples demonstrating the expected input–output format, a predefined sequence of reasoning steps that guides the model to analyze narrative coherence and clinically relevant linguistic cues and a strict output constraint requesting only the final class label. Implicit pattern matching is ensured by the embedded examples whereas the stepwise guidance limits how the model works with the transcript, making it less variable in output and promoting some form of consistency in prediction. The training model is used to pass each transcript the generated text and autoregressive decoding are analyzed with the help of simple rule-based predictions. The rules to retrieve the predicted label, no fine-tuning, no confidence calibration or post processing applied. This prompt driven design enables controlled, interpretable inference and proved particularly effective for GPT-OSS under structured reasoning constraints.

This design enables a controlled evaluation of LLM reasoning capabilities on clinical transcripts while minimizing overfitting and data leakage. By relying exclusively on text-only, prompt-based inference and emphasizing guided fewshot prompting as the central reasoning mechanism, the framework supports a fair comparison across different LLM architectures and highlights the strong effectiveness of GPT-OSS under structured reasoning constraints.

4.4.3 Vision Language Model and LLM+VLM

The VLM-based approach for dementia diagnosis uses three main prompting strategies. In zeroshot prompting, the model is provided only with instructions on how to classify a patient's description, such as the Cookie Theft picture and predicts probableAD or control without any examples or intermediate reasoning. In fewshot prompting, the model receives a small set of labeled patient examples alongside the

instructions, allowing it to implicitly compare the new patient's description to prior cases and generate a final classification but without explicitly verbalizing its reasoning. In Chain-of-Thought (CoT) prompting, the model is guided to reason step by step analyzing key elements mentioned in the description assessing linguistic quality and integrating evidence before producing a diagnosis. This methodology replicates the line of logic of the model which generates a systematic description and the last category, increasing transparency and interpretability.

The VLM+LLM model employs superior prompting methods to diagnose dementia, first a Vision Language Model analyses all visible objects, people and activities to form structured ground truth. The scenes are classified into People and Actions (identification of each individual, actions, body posture and gestures), Danger Cues (Critical) (e.g., tipping stool, overflowing water, unsafe exposures) and Objects and Locations (the items of the kitchen, the furniture, the windows, the doors, details of the surrounding). Such factual list is the basis of diagnosis.

Three prompting strategies guide subsequent reasoning by LLMs. Guided-only uses the four-step framework content mapping, linguistic biomarker detection, cognitive integration and final diagnosis—without examples enforcing stepwise reasoning via detailed sub-questions and capturing intermediate reasoning in the output JSON. Guided + FewShot combines the same structured guidance with labeled Control and ProbableAD patient examples allowing implicit pattern matching to produce final diagnostic JSON without showing reasoning steps. Guided Chain-of-Thought (CoT) also uses the structured framework without examples but requires explicit step-by-step reasoning at each stage, making the process transparent and auditable.

Chapter 5

Result Analysis

This section demonstrates analysis of the proposed design solutions, including performance metrics for dementia label classification tasks. In addition, performance evaluation is demonstrated across different modalities through structured data tables followed by detailed result analysis.

5.1 Analysis of Design Solutions

Model performance for the dementia classification task was evaluated using accuracy,area under the ROC curve (AUC), precision, F1-score and recall.

Classification Metrics:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

While accuracy measures the total number of correct predictions, it can give a false impression of performance on imbalanced datasets, where the model might predict only the majority class.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision indicates how many of the predicted positive cases are actually correct and is particularly important in medical diagnosis, where false positives can lead to unnecessary treatment.

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall indicates the model's effectiveness in capturing all actual positive cases, a critical factor in clinical screening where overlooking positives can have serious consequences.

F1-score:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score balances precision and recall and is especially suitable for imbalanced datasets.

AUC (Area Under the ROC Curve):
AUC measures the discriminative ability of a classifier across all threshold levels. A value of 1.0 indicates perfect separation, while 0.5 represents random guessing.
The false positive rate (FPR) is defined as:

$$\text{FPR} = \frac{FP}{FP + TN}$$

5.2 Performance Evaluation

5.2.1 CLAP:

The winning configurations for the top models were: Random Forest with SelectKBest feature selection (30 features) and PCA (0.95 Text / 0.90 Audio) when max pooling was used for audio segments whereas LightGBM with RFE (20 features) and PCA (0.95/0.95) under mean pooling for audio.These setups achieved the highest accuracy and AUC for each model and formed the basis for subsequent analyses.

Model	Combination	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	**Text+Audio+Features**	**0.9083**	**0.9182**	**0.9083**	**0.9083**	**0.9478**
	Text+Audio	0.6788	0.6882	0.6788	0.6789	0.7586
	Audio+Features	0.8532	0.8733	0.8522	0.8527	0.9315
	Text+Features	0.8756	0.8861	0.8756	0.8714	0.9396
	Features_Only	0.87156	0.8996	0.87156	0.8708	0.9498
	Text_Only	0.6789	0.6882	0.6789	0.6789	0.7149
	Audio_Only	0.6330	0.6404	0.6330	0.6333	0.6905
LightGBM	**Text+Audio+Features**	**0.8991**	**0.9173**	**0.8991**	**0.8989**	**0.9675**
	Text+Features	0.8899	0.9112	0.8899	0.8896	0.9607
	Audio+Features	0.87255	0.8997	0.87255	0.8708	0.9668
	Text+Audio	0.6147	0.62183	0.6147	0.61493	0.6573
	Features_Only	0.87156	0.8997	0.87156	0.8708	0.9512
	Text_Only	0.6330	0.6432	0.6330	0.6327	0.7031
	Audio_Only	0.6055	0.6196	0.6055	0.60377	0.6814

Table 5.1: Feature Ablation Study, impact of features on model performance.

Across both models, the Text + Audio + Features configuration consistently achieves the best performance, confirming the complementary nature of linguistic and acoustic features. Random Forest attains an accuracy of 90.83% and an F1-score of 0.9083, while LightGBM achieves comparable performance with 89.91% accuracy and 0.8989 F1-score, along with the highest AUC (0.9675). These results demonstrate that multimodal fusion significantly enhances discriminative capability compared to unimodal or partial combinations.

Text + Features and Audio + Features have the highest performance compared to Text+ Audio which suggests that handcrafted features are more reliable and informative signals when using in conjunction with text rather than raw audio. It is important to note that in unimodal experiments, Features-only configurations also perform significantly better than any of the other configurations (Text-only and Audio-only) based on all measures which indicates the high diagnostic quality of engineered features. Text-only models achieve moderate performance while Audio-only settings produce the weakest results, reflecting the inherent variability

and sparsity of acoustic cues in dementia speech data.

Altogether, this ablation experiment proves that multimodal integration is important to powerful dementia detection where handcrafted features are central. The similar trends of both Random Forest and LightGBM also confirm the stability and universalizability of the results.

Model	Selection Method	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	**SelectKBest**	**0.8991**	0.9068	**0.8991**	**0.8992**	0.9464
	RFE	0.8899	**0.9048**	0.8898	0.8898	**0.9491**
	SelectFromModel	0.8715	0.8861	0.8716	0.8714	0.9600
	None	0.8532	0.8733	0.8532	0.8527	0.9583
LightGBM	**RFE**	**0.8991**	**0.9173**	**0.8991**	**0.8989**	0.9675
	SelectKBest	0.8899	0.9112	0.8899	0.8896	**0.9687**
	SelectFromModel	**0.8890**	**0.9113**	**0.8991**	**0.8895**	0.9647
	None	0.8807	0.9053	0.8807	0.8802	0.9559

Table 5.2: Feature Selection Ablation Metrics for Top Models

The results of this table can be summarized as follows: compared to a noselection baseline of Random Forest and LightGBM classifiers, SelectKBest, Recursive Feature Elimination (RFE) and SelectFromModel methods have a greater impact. In the case of Random Forest, all feature selection measures are better than the baseline which is that discarding redundant and noisy features will increase the generalization. SelectKBest and RFE have the highest accuracy (89.91) and good F1-scores whereas SelectFromModel has slightly lower accuracy but still better results in comparison to the no-selection setup which shows the weakest results.

In the case of LightGBM, feature selection consistently and stabilizes improvement in all methods. RFE has the highest overall performance, accuracy (89.91%), F1-score (0.8898) and AUC (0.9675) and proves to be useful in gradient-boosted models. SelectKBest is equally competitive and has the best AUC (0.9687), whereas SelectFromModel is not inferior. Although there are no feature selection, like in Random Forest, the performance is decreased.

These findings, in general, demonstrate the importance of feature selection in enhancing robustness and generalization in high dimensional multimodal spaces of features. The stability of the gains in both classifiers shows a lesser overfitting and better decision boundary. Among the evaluated methods, RFE emerges as the most reliable and stable, while SelectKBest offers a computationally efficient alternative with comparable performance.

Classifier	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.8807	0.8928	0.8807	0.8703	0.9437
LightGBM	**0.8991**	**0.9173**	**0.8991**	**0.8989**	**0.9675**
LDA	0.8899	0.8984	0.8899	0.8805	0.9664
Naive Bayes	0.8807	0.8848	0.8807	0.8809	0.9583
XGBoost	0.8807	0.8928	0.8807	0.8802	0.9437
Logistic Regression	0.8257	0.8370	0.8267	0.8258	0.9037
SVM (Linear)	0.6789	0.7066	0.6789	0.6571	0.8184
SVM (RBF)	0.8165	0.8255	0.8165	0.8166	0.9050
KNN	0.6605	0.7247	0.6605	0.6462	0.7325

Table 5.3: Full Metrics for Best Configuration of Each Model

This table summarizes the performance of multiple classical classifiers using the final selected feature set. Tree-based ensemble models achieve the strongest results overall. LightGBM performs best across nearly all metrics, with the highest accuracy (89.91%), precision (0.9173), F1-score (0.8989) and AUC (0.9675), indicating superior discriminative power. Random Forest also performs strongly, achieving comparable accuracy (88.07%) with a balanced precision–recall trade-off.

LDA and Naive Bayes achieve competitive performance suggesting good class separability and approximate distributional assumptions in the selected features. XGBoost performs similarly to Random Forest but does not surpass LightGBM indicating limited additional benefit from boosting in this feature space.

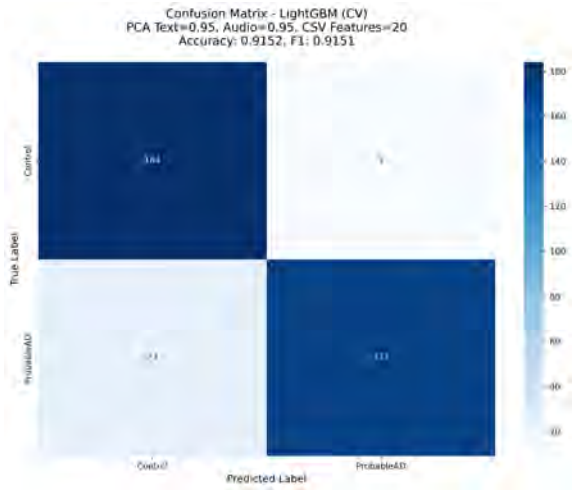


Figure 5.1: LightGBM confusion matrix on cross-validation set

In contrast, SVMs, particularly the linear variant, show reduced performance implying that the decision boundary is not well captured by linear models. Logistic Regression yields moderate results but lags behind ensemble and probabilistic approaches. KNN performs the worst across all metrics reflecting its sensitivity to noise and feature scaling in high-dimensional data.

Comparison With Existing Work

Compared with recent multimodal dementia detection studies, the proposed framework demonstrates consistently superior performance on the Pitt Corpus. Mehdoui et al.[55] combine large pretrained language models (GPT BERT) with CLAP audio embeddings using context-based fusion and cross-attention, reporting a best accuracy and F1-score of 83.33% on the Pitt Corpus but they also included MCI in their dataset which made 309 dementia samples and 243 control samples.

While their approach highlights the benefit of contextual multimodal embeddings, it relies on complex inference pipelines and large-model reasoning. In contrast, our proposed method achieves markedly higher performance, reaching 89.53% accuracy and 0.955 AUC, demonstrating stronger discriminative capability while maintaining a more structured and interpretable learning framework. Importantly, these results were obtained using the same train-test split and the same number of dementia and control samples as this GPT based study ensuring a fair comparison.

Study	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Mehdoui et al. (2023)	0.8333	0.8334	0.8333	0.8333
Our method	**0.8953**	**0.8971**	**0.8953**	**0.8951**
Shakeri et al. (2024)	0.9000	0.9500	0.8400	0.8900
Our method	**0.9461**	**0.9521**	**0.9461**	**0.9463**
Ortiz et al. (2023)	0.9036	0.9111	0.9111	0.9111
Our method	**0.9390**	**0.9396**	**0.9390**	**0.9391**

Table 5.4: Result Comparison with Existing Work

Shakeri et al.[47] evaluated classical machine learning and transfer learning approaches for dementia detection using speech transcripts from the DementiaBank Pitt Corpus, focusing primarily on linguistic features extracted from the Cookie Theft task. Using a balanced subset of the Pitt Corpus (243 Control and 255 AD samples) and 10-fold cross-validation their best-performing classical model (Random Forest) achieved 90.00% accuracy, with an F1-score of 0.89. While this result demonstrates the effectiveness of carefully engineered linguistic features the approach remains unimodal and relies solely on text-based representations.

In contrast, under the same Pitt Corpus setting and equivalent class balance, our proposed multimodal framework achieves 94.61% accuracy and an F1-score of 0.9463, substantially outperforming their results. This improvement highlights the advantage of integrating multimodal information and structured feature fusion over text-only pipelines. Despite using comparable data splits, our method delivers stronger discriminative performance while maintaining model interpretability, confirming the benefit of multimodal learning for robust dementia detection on the Pitt Corpus.

Ortiz et al. [27] proposed a deep learning–based multimodal architecture combining text and audio features for dementia detection on the DementiaBank Pitt Corpus. Their best-performing configuration, which used BERT embeddings with a BiLSTM-

based text model, achieved an accuracy of 90.36% and an F1-score of 0.9111. Although their work demonstrates the potential of deep multimodal architectures, it requires complex neural designs, extensive training, and higher computational overhead, which may limit scalability and interpretability in clinical settings.

By comparison, using the same Pitt Corpus distribution and maintaining a consistent split ratio, our approach achieves 93.90% accuracy with an F1-score of 0.9391, outperforming them while relying on a more structured and computationally efficient framework. Rather than deep end-to-end fusion, our method leverages contrastive multimodal representations and carefully selected handcrafted features resulting in superior performance with improved transparency. These results demonstrate that high diagnostic accuracy on the Pitt Corpus can be achieved without heavily parameterized deep architectures, making the proposed framework more practical for real-world dementia screening.

5.2.2 Large Language Models(LLM's)

Model	**Prompting Strategy**	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**
Qwen3-4B	Zeroshot	44.58	0.570	0.447	0.496
	Fewshot	57.23	0.593	0.566	0.535
	CoT	54.02	0.660	0.529	0.414
	Guided	54.82	0.720	0.537	0.421
	Guided + Fewshot	**62.05**	**0.679**	**0.620**	**0.581**
Mistral	Zeroshot	52.61	0.760	0.514	0.370
	Fewshot	53.41	0.607	0.524	0.413
	CoT	53.41	0.611	0.530	0.540
	Guided	58.23	0.586	0.579	0.572
	Guided + Fewshot	**58.63**	**0.600**	**0.582**	**0.569**
GPT-OSS	Zeroshot	55.82	0.611	0.549	0.481
	Fewshot	63.86	0.655	0.634	0.624
	CoT	63.45	0.636	0.634	0.635
	Guided	66.87	0.681	0.672	0.665
	Guided + Fewshot	**68.07**	**0.698**	**0.681**	**0.676**

Table 5.5: Classification performance of Large Language Models (LLMs).

The results demonstrate that prompting strategy has a substantial impact on the performance of Large Language Models (LLMs). Across all evaluated models, Guided + Fewshot prompting consistently achieved the highest accuracy and F1-scores, indicating improved classification stability and reliability. Guided prompting provides structured reasoning steps that help the models focus on clinically relevant linguistic cues, while the addition of few-shot examples further enhances contextual understanding. Among the models, GPT-OSS achieved the best overall performance with an accuracy of 68.07%, followed by Qwen3-4B. Notably, Qwen3-4B showed a significant improvement from its zeroshot accuracy of 44.58% to 62.05% with Guided + Fewshot prompting, highlighting the effectiveness of structured guidance in boosting model capability. Although Mistral demonstrated comparatively smaller gains, guided prompting still produced its strongest results, reinforcing the importance of prompt design. Overall, these findings suggest that structured guidance combined with example-based prompting leads to more stable, interpretable, and

reliable LLM-based dementia classification, emphasizing that prompt engineering plays a critical role in maximizing model performance.

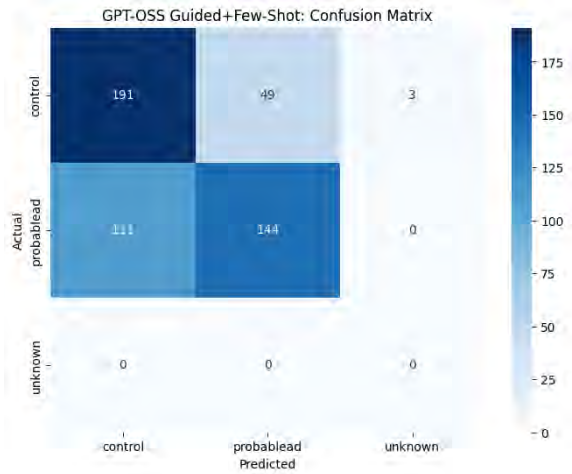


Figure 5.2: Confusion Matrix of GPT-OSS (Guided + Fewshot) Prompting

5.2.3 Vision Language Models(VLM's)

Model	Prompting Strategy	Accuracy (%)	Precision	Recall	F1-score
LLaVA-v1.6 (Vicuna-7B-hf)	Zeroshot	55.22	0.556	0.642	0.596
	Fewshot	52.61	0.522	0.940	0.671
	CoT	50.00	0.513	0.806	0.627
Qwen2.5-VL (3B-Instruct)	**Zeroshot**	**56.83**	**0.561**	**0.722**	**0.631**
	Fewshot	54.22	0.532	0.521	0.674
	CoT	48.99	0.504	0.815	0.623
Qwen2.5-VL+ Qwen2.5-Instruct	Guided	60.44	0.591	0.737	0.656
	Guided +Few shot	61.53	0.802	0.302	0.439
	Guided CoT	53.94	0.529	0.905	0.682
LLaVA + Qwen2.5-Instruct	Guided	56.14	0.564	0.626	0.593
	Guided +Few shot	**62.05**	**0.659**	**0.537**	**0.592**
	Guided CoT	56.16	0.546	0.909	0.683

Table 5.6: Classification performance of Vision–Language Models (VLMs).

The performance evaluation of Vision–Language Models (VLMs) shows that prompting strategies significantly affect classification outcomes. Zeroshot prompting provides a moderate baseline—for example, LLaVA-v1.6 achieved 55.22% accuracy with a 0.596 F1-score and Qwen2.5-VL (3B-Instruct) reached 56.83% accuracy and 0.631 F1-score. Fewshot prompting generally improves recall, as seen with LLaVA-v1.6 (recall 0.940) and Qwen2.5-VL (recall 0.521), though it may trade off precision. Guided prompting without examples enhanced overall accuracy for hybrid models, such as Qwen2.5-VL + Qwen2.5-Instruct (60.44% accuracy, 0.656 F1) and LLaVA + Qwen2.5-Instruct (56.14% accuracy, 0.593 F1). Guided + FewShot sometimes boosts precision (e.g., LLaVA + Qwen2.5-Instruct, 0.659 precision) but can reduce recall. Chain-of-Thought (CoT) prompting emphasizes structured reasoning, achieving high recall in hybrid models (e.g., LLaVA + Qwen2.5-Instruct recall 0.909, F1 0.683), demonstrating the benefit of explicit stepwise analysis. Overall, these results

indicate that the choice of prompting strategy has a greater impact than model size on VLM-based dementia classification performance.

5.3 Discussion

Some prior studies use textual data augmentation such as synonym replacement to increase training data. However, in dementia detection, such techniques risk distorting clinically meaningful linguistic patterns. Speech from individuals with dementia is often marked by reduced lexical diversity, word-finding difficulties and reliance on high-frequency vocabulary. Replacing words with semantically similar but more complex synonyms may introduce artificial patterns causing models to learn augmentation artifacts rather than genuine cognitive markers. To preserve linguistic authenticity, data augmentation was not applied in this study.

The experimental results reveal clear trends across classifiers and feature combinations. K-Nearest Neighbors (KNN) overfit the training data but performed poorly on the test set (66.05% accuracy, F1 0.6462, AUC 0.7325), showing poor generalization due to sensitivity to noise and inability to capture complex feature interactions. SVM with an RBF kernel underfit audio features while linear SVM performed slightly better (67.89% accuracy, F1 0.6571, AUC 0.8184), indicating many features are largely linearly separable but non-linear kernels struggled with limited audio-only data. Logistic Regression achieved moderate performance (82.57% accuracy because linear models capture only straightforward relationships. Ensemble and tree-based models, including Random Forest performed strongly by capturing interactions among features and reducing overfitting. Feature selection was critical: audio-only features underperformed for non-linear models. Overall, ensemble and tree-based classifiers outperform linear or kernel-based models, multimodal integration improves results and our LightGBM framework delivers state-of-the-art performance with a simpler, interpretable and efficient architecture compared to complex attention-heavy or large-model pipelines.

The chain-of-thought method reveals that there are a number of reasoning related problems that complicate output parsing. This is despite being told in no uncertain terms to provide its reasoning in plain text. special <think> tags, which in turn should be eliminated in post-processing. Such conduct demonstrates that prompt-level instructions cannot be effective when patterns acquired during pre-training or alignment are superior. Consequently, it is hard to implement strict output formats by prompting, especially when models produce internal reasoning in forms other than those desired.

These problems are not entirely solved by making the prompts more complex and allowing longer outputs. Despite the idea that the upper token cap was slowly raised up to 512 tokens in the zeroshot environment to 1024 in the chain-of-thought environment, there were still errors and as many as 32,768 tokens in the fewshot configuration, longer answers tend to add more variability, as opposed to better structure consistency. Also, the model does not often result in clear statements of confidence, which causes a lot of Unknowns. This difficulty in expressing uncertainty is especially problematic for clinical decision-support systems, where well-calibrated

confidence is essential for building trust and supporting informed decisions.

Using a complex guided prompt, GPT-OSS-20B achieved the highest accuracy of 69.67%, demonstrating that larger and more instruction-robust models can effectively follow long, structured prompts. In contrast, applying the same prompt to smaller models such as Mistral and Qwen increased Unknown predictions rather than improving accuracy, indicating limitations in handling complex prompt structures. GPT-OSS-20B's superior multi-step reasoning and long-context instruction retention enabled effective processing of the 1,500-word guided prompt. Interestingly Qwen3-4B outperformed the larger Mistral-7B in fewshot and chain-of-thought settings, demonstrating that parameter count alone does not determine task performance. This advantage stems from architectural and instruction tuning differences, as Qwen3 emphasizes structured reasoning, while Mistral is optimized for fluent conversational generation. Additionally, Mistral's sliding window attention can discard early fewshot examples and exhibits lower formatting conformity (75%) compared to Qwen3 (95%), leading to parsing failures. These results highlight that task specific instruction tuning is more critical than model size for reliable clinical classification.

Our experiments show that vision-language model performance is strongly shaped by architectural design and prompting strategy. For standalone VLMs like LLaVA-v1.6 and Qwen2.5-VL, zeroshot prompting consistently outperforms fewshot and chain-of-thought (CoT) approaches. This is due to attention dilution: adding fewshot examples or CoT reasoning increases token context reducing the attention allocated to critical visual features. As a result, fewshot prompting inflates recall but lowers precision while CoT often approaches random performance because small models cannot maintain both visual representations and multi step reasoning simultaneously.

In hybrid approach,Chain-of-thought prompting asks the LLM to remember and process many more pieces of information using its own intermediate judgments, reasoning explanations and task rules. By the time it reaches a final decision, the model is relying more on its own generated labels than the actual evidence. This results in high recall as it catches obvious cases but very poor precision mislabeling borderline cases lowering overall accuracy. Architecturally, this happens because attention in transformer models is limited. In CoT, most of the model's attention goes to recent outputs while the original facts get very little focus.

Moreover, preliminary fine tuning experiments with the Mistral model were performed, but the limited time frame and computational resources constrained further optimization.

Chapter 6

Conclusion

Here we conclude our research by stating the summary of findings, contribution to the field, recommendations for future work and finally, limiations of this work.

6.1 Summary of Findings

6.1.1 CLAP

Across all experiments, multimodal fusion consistently outperforms unimodal and partial configurations confirming that linguistic and acoustic features provide complementary diagnostic information. In particular, the Text + Audio + Features setup achieves the strongest and most stable performance for both Random Forest and LightGBM with LightGBM yielding the best overall results including 89.91% accuracy, 0.8989 F1-score and the highest AUC of 0.9675.

The ablation results show that handcrafted features are especially important. Models using these features perform better than text-only or audio-only models and even better than using text and audio together without features. Audio-only models perform the worst likely due to noise and variability in speech signals. Feature selection further improves performance. Methods like RFE and SelectKBest consistently outperform models without feature selection, helping reduce noise and improve generalization. Among all classifiers, tree-based models perform bes, while simpler models like KNN and linear SVM lag behind.

Compared to prior work, the proposed approach achieves higher accuracy and AUC on the Pitt Corpus while using simpler and more interpretable models making it a practical and reliable solution for dementia detection.

6.1.2 Large Language Model

Most previous investigations on dementia detection using language models have relied on supervised learning or fine-tuning. While these approaches are highly effective, they necessitate significant computational resources and labeled data. Recent assessments of LLMs in zeroshot or basic chain-of-thought (CoT) have reported limited performance, finding that unadapted LLMs are not yet reliable for clinical diagnosis [30]. This reveals a gap concerning the feasibility of dementia detection

using freely accessible, training-free models.

Our study evaluates dementia label detection using free and open-access LLMs, without fine-tuning or parameter customization. Despite restrictions, the Guided + Fewshot prompting technique beats zeroshot, fewshot, CoT and guided approaches in all models, with up to 68.07% accuracy and 0.695 F1-score. These findings show that organized guidance and exemplar-based prompting can significantly improve reasoning performance in resource-limited environments.

6.1.3 Vision Language Model

Our work does not entail fine-tuning counting of explicit cues and threshold-related rules but rather adopts prompt-driven reasoning procedures that favor more intricate interpretation of visual and linguistic data. In experiments, standalone VLMs perform moderately, whereas always when combined with LLM reasoning and VLM perception, the performance is improved. It has the highest overall outcome with best accuracy and F1-score of 62.05 percent and 0.592 respectively, beating the best standalone result of VLM (Qwen2.5-VL 7B) with highest accuracy and F1 of 60.44 percent and 0.682 respectively.

It is interesting to note that standalone VLMs tend to perform poorly with Fewshot and CoT prompting as suggesting the inability to process complex logical instructions. Conversely, the hybrid framework has far fewer cases of poor prompting with the accuracy boosting to 62.05%. The hybrid pipeline by separating the visual perception and logical reasoning, takes advantage of each part in a better way compared to the monolithic VLMs. In general, these results confirm the idea that the combination of perception and reasoning is useful in dementia assessment even in a completely training free, prompt-based environment.

6.2 Contributions to the Field

The work is relevant to research on dementia detection because we examine the ability of foundation models in a fully training-free environment. The work compares Large language Models (LLM), Vision-Language Models (VLM) and audio-text representations with CLAP, showing that it is possible to do meaningful cognitive assessment even without fine-tuning.

The experimental results demonstrate that simple CLAP embeddings are competitively effective which implies the usefulness of audio-text alignment models in the process of retrieving acoustic patterns associated with cognitive deterioration. This finding supports the claim that speech-based dementia analysis can be effectively based on lightweight use of CLAP without involving the need to elaborate on feature engineering or multimodal fusion techniques.

Besides that the study provides a thorough study of several prompting strategies such as the zeroshot, fewshot chain-of-thought and guided prompting showing that structured prompts are of great importance in increasing reasoning consistency and

classification performance. The findings also confirm that the combination of language thinking with visual and auditory data leading to the enhanced performance compared to mono-modes which confirms the complementary functions of the text, vision and audio modalities.

Generally, this paper provides a cost-effective, reproducible and scalable framework of assessing dementia-related issues based on free and open-access models to provide practical understanding of future studies on prompt-driven multimodal reasoning in healthcare systems.

6.3 Recommendations for Future Work

Although the present study concentrates on the inference of prompts by using free and open-source models, the future research can further develop this study in various ways. Among the important measures is to explore the effect of fine-tuning techniques such as parameter-efficiency to compare the effectiveness of task-specific adaptation and prompt-only methods.

The other potential avenue is the enhanced use of audio-text models like CLAP. Here, CLAP was applied in a simple embedding-style fashion, but it can be considered in future research that more sophisticated fusion techniques, fine-tuning or prompt-guided alignment can be used to learn more subtle acoustic features related to cognitive impairment. A more efficient combination of speech, text and image modalities would give a more lucid account of cognitive behavior and enhance the general assessment of dementia.

Also, further research can be conducted on the effectiveness of more sophisticated commercial models and developing a web-based intelligent agent with the ability to analyze in an interactive and real-time fashion. These extensions would increase usability, scalability and realistic deployment, which would make the proposed framework stronger in a real-world environment.

6.4 Limitation

Although the positive outcomes were achieved during this research, there are a number of limitations that ought to be acknowledged. The limited data size is one of the major constraints since it might also influence the external validity of the results. Though critical consideration was made, greater and less homogenous datasets would give more robust statistical support and wider range of variation in cognition.

The other limitation is the issue of the computational resources and time. While it used an RTX 5090 GPU and this high-end hardware was only accessible to this for a limited period. This limited the extent of experimentation especially on very high fine-tuning scale, protracted hyperparameter exploration and more multimodal fusion approaches.

Moreover, the complicated set up of the experimental environment was an added problem challenges. Depending on large-scale language, vision-language models took too much time to configure dependencies and this shortened time allocated to large-scale tasks testing. Consequently, the research gave more importance to stable and repeatable cases throughout more expansive architectural discovery.

These constraints represent real world difficulties that are usually faced in practice test environments and offer a strong incentive to work on larger datasets in the future shorter computational latency and smoother pipelines of experiment.

6.5 Conclusion

The free dementia detection can be trained successfully by taking advantage of multimodal representations, contrastive learning and prompt driven reasoning with large foundation models. Results on the Pitt Corpus show that CLAP based audio text alignment delivers the strongest performance, emphasizing the value of combining acoustic and linguistic cues to capture dementia related speech characteristics. Experiments with large language and vision language models further demonstrate that guided and fewshot prompting significantly enhances inference quality, allowing models to identify clinically meaningful patterns without parameter finetuning. Overall, these findings highlight the potential of prompt based multimodal frameworks as scalable, cost efficient alternatives to traditional supervised methods for early dementia screening.

Bibliography

[1] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. DOI: 10.1080/00401706.1970.10488634.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. DOI: 10.1007/BF00994018.

[3] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. DOI: 10.1109/78.650093.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[5] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004. DOI: 10.1023/B:STCO.0000035301.49549.88.

[6] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Critical Care*, vol. 9, no. 1, pp. 112–118, 2005. DOI: 10.1186/cc3045.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint*, 2014. arXiv: 1409.0473.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 1, pp. 1–19, 2016. DOI: 10.1145/2939672.2939785.

[9] 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. DOI: 10.3233/JAD-150520. [Online]. Available: https://doi.org/10.3233/JAD-150520.

[10] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," in *Proceedings of Interspeech 2018*, 2018, pp. 1691–1695. DOI: 10.21437/Interspeech.2018-2278. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-2278.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, 2019. arXiv: 1810.04805.

[12] Y. Liu, M. Ott, and N. e. a. Goyal, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint*, 2019. arXiv: 1907.11692.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper, and lighter," *arXiv preprint*, 2019. arXiv: 1910.01108.

[14] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad, and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint*, 2020. arXiv: 2006.11477.

[16] T. Brown, B. Mann, and N. e. a. Ryder, "Language models are few-shot learners," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. eprint: 2005.14165.

[17] Z. Lan, M. Chen, and S. e. a. Goodman, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint*, 2020. arXiv: 1909.11942.

[18] J. Lee, W. Yoon, and S. e. a. Kim, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. DOI: 10.1093/bioinformatics/btz682.

[19] J. Chen, Y. Wu, X. Zhang, and B. Liu, "Slime: Self-supervised learning-based interpretability method for deep neural networks," in *NeurIPS*, 2021.

[20] F. Agbavor and H. Liang, "Predicting dementia from spontaneous speech using large language models," *PLOS digital health*, vol. 1, no. 12, e0000168, 2022.

[21] A. Chowdhery, S. Narang, and J. e. a. Devlin, "Palm: Scaling language modeling with pathways," *Google AI Blog*, 2022.

[22] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, *Clap: Learning audio concepts from natural language supervision*, 2022. arXiv: 2206.04769 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2206.04769.

[23] J. Wei, X. Wang, and D. e. a. Schuurmans, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint*, 2022. arXiv: 2201.11903.

[24] A. Q. Jiang et al., *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, *Visual instruction tuning*, 2023. arXiv: 2304.08485 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.08485.

[26] OpenAI, "Gpt-4 technical report," *OpenAI Research*, 2023. [Online]. Available: https://openai.com/research/gpt-4.

[27] 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. 126 413, 2023.

[28] K. Panesar and M. B. P. C. de Alba, "Natural language processing-driven framework for the early detection of language and cognitive decline," *Language and Health*, vol. 1, no. 2, pp. 20–35, 2023.

[29] H. Touvron, L. Martin, and K. e. a. Stone, "Llama 2: Open foundation and fine-tuned chat models," *Meta AI*, 2023.

[30] Z. Wang et al., "Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today," *arXiv preprint arXiv:2306.01499*, 2023.

[31] E. Ambrosini et al., "Automatic spontaneous speech analysis for the detection of cognitive functional decline in older adults: Multilanguage cross-sectional study," *JMIR aging*, vol. 7, e50537, 2024.

[32] F. de Arriba-Pérez and S. García-Méndez, "Leveraging large language models through natural language processing to provide interpretable machine learning predictions of mental deterioration in real time," *Arabian Journal for Science and Engineering*, pp. 1–15, 2024.

[33] F. de Arriba-Pérez, S. García-Méndez, J. Otero-Mosquera, and F. J. González-Castaño, "Explainable cognitive decline detection in free dialogues with a machine learning approach based on pre-trained large language models," *Applied Intelligence*, vol. 54, no. 24, pp. 12 613–12 628, 2024.

[34] J. Banks and T. Warkentin, *Gemma: Introducing new state-of-the-art open models*, https://ai.google.dev/gemma, Published Feb 21, 2024; accessed Jun 18, 2025, Feb. 2024.

[35] C.-P. Chen and J.-L. Li, "Profiling patient transcript using large language model reasoning augmentation for alzheimer's disease detection," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2024, pp. 1–4.

[36] J.-M. Chen et al., "Performance assessment of chatgpt vs bard in detecting alzheimer's dementia," *arXiv preprint arXiv:2402.01751*, 2024.

[37] X. Du et al., "Enhancing early detection of cognitive decline in the elderly: A comparative study utilizing large language models in clinical notes," *EBioMedicine*, vol. 109, 2024.

[38] S. A. Freja and Y. Hallaj, "Early dementia detection in speech transcripts using machine learning and large language models," M.S. thesis, UIS, 2024.

[39] S. S. Huang et al., "Fact check: Assessing the response of chatgpt to alzheimer's disease myths," *Journal of the American Medical Directors Association*, vol. 25, no. 10, p. 105 178, 2024.

[40] K. Lin and P. Y. Washington, "Multimodal deep learning for dementia classification using text and audio," *Scientific Reports*, vol. 14, no. 1, p. 13 887, 2024.

[41] U. Lokala, R. H. Desai, V. L. Shalin, and A. Sheth, "Artificial intelligence based detection of early cognitive impairment using language, speech, and demographic features: Model development and validation," 2024.

[42] Meta AI Team, *Introducing meta llama 3: The most capable openly available llm*, https://ai.meta.com/blog/meta-llama-3/, Published April 18, 2024; accessed June 18, 2025, Apr. 2024.

[43] Microsoft, *Phi-3-Mini-4K-Instruct: a 3.8B-parameter lightweight instruction-tuned language model*, https://huggingface.co/microsoft/Phi-3-mini-4k-instruct, Accessed: 2025-06-18; MIT license, Sep. 2024.

[44] T. Mo, J. Lam, V. Li, and L. Cheung, "Leveraging large language models for identifying interpretable linguistic markers and enhancing alzheimer's disease diagnostics," *medRxiv*, pp. 2024–08, 2024.

[45] F. F. Poor, H. H. Dodge, and M. H. Mahoor, "A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language and vision," *Computers in Biology and Medicine*, vol. 182, p. 109 199, 2024.

[46] M. Ribeiro et al., "A methodology for explainable large language models with integrated gradients and linguistic analysis in text classification," *arXiv preprint arXiv:2410.00250*, 2024.

[47] A. Shakeri, S. A. Freja, Y. Hallaj, and M. Farmanbar, "Uncovering linguistic patterns: A machine learning exploration for early dementia detection in speech transcripts," in *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*, IEEE, 2024, pp. 1–8.

[48] T. Shang et al., "Leveraging social determinants of health in alzheimer's research using llm-augmented literature mining and knowledge graphs," *arXiv preprint arXiv:2410.09080*, 2024.

[49] R. Soleimani, S. Guo, K. L. Haley, A. Jacks, and E. Lobaton, "Dementia detection by in-text pause encoding," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2024, pp. 1–4.

[50] S. Suravee, T. Stoev, S. Konow, and K. Yordanova, "Assessing large language models for annotating data in dementia-related texts: A comparative study with human annotators," in *INFORMATIK 2024*, Gesellschaft für Informatik eV, 2024, pp. 487–498.

[51] J. Wang et al., "Augmented risk prediction for the onset of alzheimer's disease from electronic health records with large language models," *arXiv preprint arXiv:2405.16413*, 2024.

[52] Alibaba Cloud Community, *Alibaba introduces qwen3, setting new benchmark in open-source ai with hybrid reasoning*, https://www.alibabacloud.com/blog/alibaba-introduces-qwen3-setting-new-benchmark-in-open-source-ai-with-hybrid-reasoning_602192, Accessed: 2025-06-18, Apr. 2025.

[53] S. Bai et al., *Qwen2.5-vl technical report*, 2025. arXiv: 2502.13923 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2502.13923.

[54] C. Li, R. Li, T. S. Field, and G. Carenini, "Delta-knn: Improving demonstration selection in in-context learning for alzheimer's disease detection," *arXiv preprint arXiv:2506.03476*, 2025. [Online]. Available: https://arxiv.org/pdf/2506.03476.

[55] S. S. Mehdoui, A. Bouzid, D. Sierra-Sosa, and A. Elmaghraby, *Dementia insights: A context-based multimodal approach*, 2025. arXiv: 2503.01226 [q-bio.NC]. [Online]. Available: https://arxiv.org/abs/2503.01226.

[56] OpenAI et al., *Gpt-oss-120b gpt-oss-20b model card*, 2025. arXiv: 2508.10925 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2508.10925.

[57] Y. Pan, B. Mirheidari, D. Blackburn, and H. Christensen, "A two-step attention-based feature combination cross-attention system for speech-based dementia detection," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[58] C. Park, A. S. G. Choi, S. Cho, and C. Kim, "Reasoning-based approach with chain-of-thought for alzheimer's detection using speech and large language models," *arXiv preprint arXiv:2506.01683*, 2025. arXiv: 2506.01683 [cs.CL].

[59] Qwen et al., *Qwen2.5 technical report*, 2025. arXiv: 2412.15115 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.15115.

[60] A. Yang et al., *Qwen3 technical report*, 2025. arXiv: 2505.09388 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2505.09388.

Appendix

.1 Prompt Template

LLM & VLM Prompts

Guided Prompt

DEMENTIA SCREENING: COOKIE-THEFT PICTURE DESCRIPTION TEST (CDPDT)
TASK: Analyze Cookie-Theft transcripts and classify as **Control** (cognitively normal) or **ProbableAD** (Alzheimer's disease). Provide a binary diagnosis + estimated MMSE score (0-30).
—
DECISION FRAMEWORK: CONTROL vs PROBABLEAD
Quick Decision Rules
CONTROL (Cognitively Normal): - Clear, organized description with logical flow - Minimal word-finding pauses (¡10%) - High semantic content (describes most key elements) - Maintains coherent narrative structure - Few repetitions or self-corrections - MMSE typically 27-30
PROBABLE AD (Alzheimer's Disease): - Fragmented or disorganized narrative - Frequent pausing/hesitations (¿15% of utterance) - Missing key story elements or relationships - Poor sequence/temporal organization - Multiple repetitions or rambling - Reduced detail and anomia (word-finding difficulty) - MMSE typically 10-24
—
KEY DISCRIMINATIVE FEATURES (RANKED BY IMPORTANCE)
1. **SEMANTIC COMPLETENESS RELEVANCE** (Most Important) Use this scoring:
Control Pattern: - Identifies core objects: mother, child/boy, girl, cookie jar, stool, water, sink, window - Describes key action: boy reaching for/taking cookies, boy on unstable stool - Notes consequences: water overflowing, mother at sink, risk of falling - 12-14 relevant units mentioned
ProbableAD Pattern: - Misses central action or relationships ("there's a boy... and a stool...") - Omits critical elements (no mention of cookie jar, or mother, or consequences) - Focuses on minor details ("the curtains are blue") while missing main events - 3-9 relevant units mentioned - May confabulate details not in picture
→ Red Flag: If ¿3 major elements (jar, boy reaching, stool, water, mother) are completely missing = strong AD indicator
—
2. **SPEECH FLUENCY PAUSING** (Very Important)
Control Pattern: - Smooth, continuous speech with minimal hesitation - Occasional natural pauses but quick recovery - ¡10% of total time spent pausing - Few "um/uh" or "like" fillers
ProbableAD Pattern: - Frequent long pauses (¿1-2 seconds) before key words - Repetitive fillers: "um... um... the... the..." - Multiple false starts: "He... she... I mean the boy..." - Visible struggle to retrieve common words ("cookie... that... you know, the sweet thing") - 20-40% of speech is pausing/hesitation
→ Red Flag: If describing simple objects requires multiple pauses and restarts = word-finding difficulty (anomia)
—
3. **NARRATIVE STRUCTURE COHERENCE** (Very Important)
Control Pattern: - Information presented in logical sequence (first describes mother, then child, then action) - Uses temporal markers: "while", "and then", "as", "because" - Connections between events are clear - Topic stays focused on picture **ProbableAD Pattern:** - Jumbled order (describes actions out of sequence) - Missing temporal/causal connectors - Topic drift ("reminds me of my own kitchen", "I had a friend...") - Contradictions or confused pronouns ("she... he... it is...") - Rambling: idea doesn't flow logically
→ Red Flag: If narrative is hard to follow or jumps around = executive dysfunction (hallmark of AD)
—
4. **REPETITION PERSEVERATION** (Important)
Control Pattern: - Mentions items once or twice - Self-corrections are rare and natural ("the boy, or actually the two children...") - ¡2 repetitions of the same word/phrase
ProbableAD Pattern: - Repeats the same word excessively: "the cookie... cookie... cookies... cookie jar..." - Repeats entire phrases: "water running, water running, the water is running..." - Perseverates on minor details: "the curtains... the curtains are open... the curtains are blowing..." - 5-10+ repetitions

→ Red Flag: Excessive repetition of single concept = semantic retrieval deficit or working memory impairment
—
5. **DETAIL ACCURACY** (Important)
Control Pattern: - Accurate object labels ("stool", "faucet", "dishcloth") - Correct relationships: "boy on stool", "mother drying dishes" - Accurate scene interpretation: "boy is reaching/stealing cookies"
ProbableAD Pattern: - Vague labels ("the thing", "stuff", "it") - Mislabeling or generic terms instead of specific objects - Incorrect relationships: "the boy is sitting" (when he's reaching), "the girl is eating" - Semantic errors: confuses who is doing what
→ Red Flag: Inability to name common objects or misunderstanding simple relationships = semantic decline
—
SCORING CHECKLIST FOR BINARY DECISION
Count how many of these abnormal indicators are present:
— Feature — Abnormal Threshold — Status — ——————————————————— — Semantic completeness — ¡9 relevant units — — — Pausing frequency — ¿15% of utterance — — — Narrative incoherence — Hard to follow flow — — — Excessive repetition — ¿4 repetitions of same element — — — Anomia/vague terms — ¿5 instances ("thing", "stuff") — — — Missing temporal markers — Uses ¡2 temporal words — — — Self-corrections — ¿3 false starts — — — Topic drift — ¿1 off-topic tangent — —
DECISION RULE: - **0-1 abnormal indicators** → **CONTROL** (MMSE: 27-30) - **2-3 abnormal indicators** → **CONTROL** (MMSE: 25-27) [borderline, slight decline] - **4-5 abnormal indicators** → **PROBABLE AD** (MMSE: 18-24) [mild-moderate AD] - **6+ abnormal indicators** → **PROBABLE AD** (MMSE: 10-17) [moderate-severe AD]
—
MMSE ESTIMATION GUIDE
MMSE 27-30 (Control): - Clear, complete description; minimal pauses; coherent; accurate - All major elements present; good temporal organization
MMSE 24-26 (Borderline/Mild): - Mostly coherent; occasional pauses; minor omissions; 1-2 vague terms
MMSE 18-23 (Mild-Moderate AD): - Visible word-finding difficulty; some disorganization; missing 2-3 key elements - Multiple pauses; repetitions present
MMSE 10-17 (Moderate-Severe AD): - Severe pausing; fragmented; multiple omissions; heavy reliance on vague terms - Poor coherence; excessive repetition; confusion about relationships
MMSE ¡10 (Severe AD): - Minimal coherent description; severe anomia; unable to maintain topic - Perseveration dominant; confabulation or hallucinations present
—
WHAT TO IGNORE (Not Relevant for Control vs AD)
- Accent, language fluency (non-native speakers can still be cognitively normal) - Background noise or transcription artifacts - Other dementia types (PPA, FTD, vascular) - focus only on AD vs Control - Personality or emotional tone - Long pauses due to thinking carefully (vs. pauses due to word-finding struggle)
—
ANALYSIS PROCESS
1. **Read the full transcript** to understand overall coherence 2. **Count relevant units** (key elements described: mother, boy, girl, jar, stool, water, sink, action, consequence) 3. **Assess pausing** (calculate approximate % of hesitation) 4. **Check narrative flow** (is it logical? are temporal markers present?) 5. **Count repetitions** (any words/phrases mentioned ¿4 times?) 6. **Evaluate word choice** (specific terms vs. vague "thing"/"stuff"?) 7. **Identify anomia indicators** (word-finding struggles, restarts, circumlocution) 8. **Apply decision rule** (count abnormal indicators; classify based on threshold) 9. **Estimate MMSE** (map severity to score range)
—
SUMMARY OF AD vs CONTROL DIFFERENTIATION
— Dimension — CONTROL — PROBABLE AD — ——————————————————— — **Overall Impression** — Organized, fluent, coherent — Fragmented, halting, disorganized — — **Key Indicator** — Describes main action clearly — Struggles to convey main action — — **Pausing** — Minimal, natural — Frequent, labored — — **Vocabulary** — Specific, accurate — Vague, generic, anomic — — **Coherence** — Logical flow, temporal markers — Jumbled, missing connectors — — **Detail Level** — 12-14 relevant units — 3-9 relevant units — — **Repetition** — ¡2 per concept — 5+ per concept — — **Typical MMSE** — 27-30 — 10-24 —
—
Apply this framework rigorously when analyzing each transcript.

Fewshot

"""Create few-shot prompt for dementia diagnosis.""" return f"""You are an expert neuropsychologist specializing in dementia assessment. Analyze the following Cookie Theft picture description transcript.
File: filename
Transcript: transcript
Task: Based on this transcript, predict:
1. **Diagnosis Label**: Choose ONE from: - ProbableAD - Control

Response Format (only this, no extra text): Label: [ProbableAD/Control] Confidence: [Low/Medium/High]"""
Filename: 002-0.txt
Transcript:
the scene is ¡in the¿ [/] in the kitchen . the mother is wiping dishes and the water is running on the floor . ¡a child is tryin(g) to get¿ [//] a boy is tryin(g) to get cookies outta [: out of] a jar and he's about to tip over on a stool . -uh the little girl is reacting to his falling . -uh it seems to be summer out . the window is open . the curtains are blowing . it must be a gentle breeze . there's grass outside in the garden . -uh mother's finished certain of the [/] the dishes . kitchen's very tidy . [+ gram] the mother seems to have nothing in the house to eat except cookies in the cookie jar . -uh the children look to be almost about the same size . perhaps they're twins . they're dressed for summer warm weather . -um you want more ? [+ exc] +¡ the mother's in a short sleeve dress . +¡ I'll hafta say it's warm . [+ exc]
Label: Control

Filename: 001-0.txt
Transcript:
mhm . [+ exc] +¡ alright . [+ exc] there's -um a young boy that's getting a cookie jar . and it [//] he's -uh in bad shape because -uh the thing is fallin(g) over . and in the picture the mother is washin(g) dishes and doesn't see it . and so ¡is the¿ [//] the water is overflowing in the sink . and the dishes might ¡get falled [* +ed] over if you don't¿ [//] fell [//] fall over there [/] there if you don't get it . and it [//] there [//] it's a picture of a kitchen window . and the curtains are very -uh distinct . but the water is +flow still flowing .
Label: ProbableAD

Filename: 034-1.txt
Transcript:
boy -uh taking cookies out of a cookie jar . [+ gram] the stool is falling . the little girl is reaching . water is running out o(f) the faucet . the water is overflowing the sink . woman is ¡washin(g) dishes¿ [//] (...) drying dishes . [+ gram] there's nothing to indicate xxx and I don't see any more action . [+ exc]
Label: Control

Filename: 216-1.txt
Transcript:
now honey I +w +l +l +ha +w had it was in the kitchen and I was the oldest ten . [+ gram] [+ exc] and if we made a mess like that you'd get a kick in the ass . [+ exc] +¡ well ‡ we have -uh spillin(g) of the water . and a kid with his cookie jar . [+ gram] and a stool is turned over . and a mother's runnin(g) the water on the floor . and what else do you want from that ? [+ exc] it +s looks like somebody's layin(g) out in the grass doesn't it ? +¡ and a kid in the cookie jar . [+ gram] and a tilted +s stool . [+ gram] what more do you want ? [+ exc] +¡ the [/] +wa the water rollin(g) on the floor . [+ gram]
Label: ProbableAD

Filename: 684-0.txt
Transcript:
-uh the boy is reaching into the cookie jar . he's falling off the stool . the little girl is reaching for a cookie . mother is drying the dishes . the sink is running over . mother's getting her feet wet . they all have shoes on . there's ¡a cup¿ [//] two cups and a saucer on the sink . the window has draw [//] withdrawn [: drawn] [* s:r] drapes . you look out on the driveway . there's kitchen cabinets . oh what's happening . [+ exc] mother is looking out the window . the girl is touching her lips . the boy is standing on his right foot . his left foot is sort_of up in the air . mother's right foot is flat on the floor and ¡her left¿ [//] she's on her left toe . -uh she's holding the dish cloth in her right hand and the plate she is drying in her left . I think I've run out of +/. [+ exc] +¡ yeah =laughs . [+ exc]
Label: Control

Filename: 711-0.txt
Transcript:
oh ‡ you want me to tell you . [+ exc] the mother and her two children . [+ gram] and the children are getting in the cookie jar . and she's doing the dishes and spilling the water . and she had the spigot on . and she didn't know it perhaps . =coughs pardon me . [+ exc] and they're looking out into the garden from the kitchen window . it's open . and the -uh cookies must be pretty good they're eating . ¡the tair [: chair] [* p:w-ret]¿ [//] -uh the chair [: stool] [* s:r] is +t -uh tilting and he's gonna fall off . and -uh ¡the lady¿ [//] the mother's splashing her shoes and stockings all up overflowing the water . and there's -um -uh a window and curtains on the window . and I can see some trees outside there . and [/] and there's dishes that +h had been washed . and she's dryin(g) them . and there's some shrub out there and +...
Label: ProbableAD

Chain-of-Thought (CoT)

You are an expert neuropsychologist specializing in dementia assessment. Analyze the following Cookie Theft picture description transcript.
Task: Analyze this transcript step by step to determine whether the speaker shows signs of dementia.
Think through the following aspects:

1. Language Fluency
Is the speech fluent or hesitant? Are there frequent pauses, fillers, restarts, or incomplete utterances?
2. Vocabulary & Word Finding
Does the speaker use specific and appropriate words, or vague expressions? Is there evidence of word-finding difficulty or circumlocution?
3. Content & Completeness
Does the speaker mention the key elements of the Cookie Theft picture (e.g., mother, children, stool, cookies, sink overflow, window)?
4. Grammar & Syntax
Are sentences grammatically well-formed, or fragmented with omissions and errors?
5. Coherence & Organization
Is the narrative logically organized and easy to follow, or scattered and disjointed?
6. Repetitions
Are there unnecessary repetitions of words, phrases, or ideas?
Step-by-step reasoning:
Provide a detailed analysis for each aspect listed above, using only the evidence present in the transcript.
Final Assessment:
Label: `ProbableAD` / `Control`
Confidence: `Low / Medium / High`

Zeroshot

"""Look at this image and analyze the description provided. Based on the characteristics of the description, classify whether it shows signs of Alzheimer's disease or is from a control (healthy) participant.
Respond in this exact format: Classification: [probablead/control]

Guided Prompt for VLM+LLM

For the VLM:

Analyze this 'Cookie Theft' image with forensic precision.
List EVERY visible object, person and action in strict categories:
1. PEOPLE & ACTIONS:
- What is each person doing?
- Body positions and gestures

2. DANGER CUES (CRITICAL):
- Is the stool tipping?
- Is water overflowing?
- Any unsafe situations?

3. OBJECTS & LOCATIONS:
- Kitchen items (plates, cups, utensils)
- Furniture
- Windows, doors, environmental details

Output a comprehensive, factual list. This is GROUND TRUTH for medical diagnosis.

For the LLM:
SYSTEM ROLE
You are a senior Neurologist specializing in Alzheimer's Disease (AD) detection. Analyze the patient's speech with OBJECTIVITY and PRECISION. Be ACCURATE - do not bias toward either Control or ProbableAD. Follow the evidence strictly.
INPUT DATA
{gt_formatted}
PATIENT TRANSCRIPT:"{patient_transcript}
ANALYSIS PROTOCOL

Compare this patient to the examples above and the visual ground truth.

STEP 1: CONTENT MAPPING (Accuracy)
- Compare the transcript to the visual reality above.
- **CRITICAL:** Did the patient miss danger cues listed above (especially stool tipping, water overflowing)?
- **CRITICAL:** Did the patient misidentify objects or actions (semantic paraphasia)?
- Check whether people, objects and actions from the lists were mentioned.

STEP 2: LINGUISTIC BIOMARKER DETECTION (Style)
- **Anomia:** Count empty or vague words ("thing", "stuff", "it", "they").
- **Fluency:** Count hesitations ("um", "uh") and sentence fragments.

- **Syntax:** Complex/coherent vs. telegraphic/broken grammar.

STEP 3: BINARY CLASSIFICATION
- **Control:** Mentions most key objects (especially danger cues), fluent speech, complex sentences (similar to Control examples above).
- **ProbableAD:** Misses danger cues from the list **or** shows frequent empty words, hesitations, or broken grammar (similar to ProbableAD examples above).

STEP 4: MMSE ESTIMATION
- Estimate the patient's MMSE score (0–30) based on comparison with example transcripts above.

FINAL OUTPUT FORMAT (Strict JSON)
{
"missed_danger_cues": ["List of danger cues from above that patient failed to mention"],

"anomia_count": (Integer),

"syntax_quality": "High"/ "Medium"/ "Low",

"clinical_score": (0–10 scale, where < 5 is Control, >= 5 is ProbableAD),

"diagnosis": "Control" or "ProbableAD",

}