CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

A Multimodal Feature Fusion and Explainable AI Driven Framework for Early Detection and Staging of Dementia

by

Muhammad Awais

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2026

CERTIFICATE OF APPROVAL

A Multimodal Feature Fusion and Explainable AI Driven Framework for Early Detection and Staging of Dementia

by

Muhammad Awais

(MCS243006)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Erum Ashraf	Bahria Uni., Islamabad
(b)	Internal Examiner	Dr. Rizwan Bin Faiz	CUST Islamabad

Dr. Sabeen Masood
Thesis Supervisor
May, 2026

Dr. M. Masroor Ahmed
Head
Dept. of Computer Science
May, 2026

Dr. M. Abdul Qadir
Dean
Faculty of Computing
May, 2026

Author's Declaration

I, **Muhammad Awais** hereby state that my MS thesis titled "**A Multimodal Feature Fusion and Explainable AI Driven Framework for Early Detection and Staging of Dementia**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Muhammad Awais)

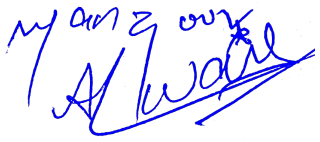
Registration No: MCS243006

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**A Multimodal Feature Fusion and Explainable AI Driven Framework for Early Detection and Staging of Dementia**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.



(Muhammad Awais)

Registration No: MCS243006

Acknowledgement

First and foremost, I express my deepest gratitude to Almighty Allah for granting me the strength, patience, and guidance to successfully complete this research work. Without His blessings and mercy, this achievement would not have been possible.

I would like to extend my sincere appreciation to my respected supervisor, Dr. Sabeen Masood, for her continuous guidance, invaluable feedback, and unwavering support throughout every stage of this thesis. Her insightful discussions, constructive criticism, and encouragement at each step played a fundamental role in shaping this research and bringing it to completion. I am truly grateful for her mentorship and dedication.

I would also like to express my heartfelt gratitude to Wahaj Ahmed/AO-SE and the Well Being Center of CUST, especially Dr. Sabahat Haqqani (Professor / HOD Psychology), for their constant moral support, motivation, and encouragement during a challenging period of my health. Their kindness, understanding, and continuous encouragement helped me stay focused and strong throughout this journey, and their support played a significant role in helping me complete this work.

I am also deeply thankful to my family for their unconditional love, prayers, and constant encouragement. Their belief in me has always been my greatest source of motivation. Finally, I would like to acknowledge my friends and colleagues for their support, discussions, and moral encouragement throughout this journey. This accomplishment would not have been possible without the support of all those who stood by me during this important phase of my academic life.

(Muhammad Awais)

Abstract

Dementia is an irreversible neurodegenerative disease, which is marked by the decrease of cognitive and linguistic abilities, and requires precision, availability, and decipherable initial diagnosis tools. Regardless of the overwhelming achievements in artificial intelligence (AI)-driven diagnostics, the available solutions possess the limitations of the poor multimodal feature integration, black-box decisions, insufficient detection at the early stage, and low generalization on heterogeneous data. This study overcomes such difficulties by suggesting a multimodal speech-based AI model that can identify stages of cognitive impairment (Healthy Control, Mild Cognitive Impairment and Dementia) without sacrificing cross-dataset robustness. The suggested system combines acoustic, linguistic, and temporal-prosodic features observed on three standardized tests of verbal fluency- Semantic Fluency Test (SFT), Category Fluency Test (CTD), and Phonemic Fluency Test (PFT) into one 300-dimensional multimodal representation of features. Classical machine learning architectures like Support Vector Machine (SVM), Random Forest (RF) and XGBoost were only applied together with such deep learning architectures like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Transformer networks to compare them comparatively in terms of discriminative performance. To achieve model interpretability, SHapley Additive exPlanations (SHAP) were used, which allows analyzing the importance of global features and explaining instances to support clinical transparency. Two datasets were evaluated experimentally to measure internal performance as well as external generalization. The most successful models in the Kaggle competition (Dementia Detection Using Speech) of multi-class: HC, MCI, Dementia have an overall accuracy of 76.19 with Dementia sensitivity of 1.000 and specificity of 0.857, which shows high screening capabilities. RF and LSTM models demonstrated 71.43% accuracy, whereas Transformer performed dismally at 54.76% which indicates that sequential modeling was the most appropriate unknown model compared to global self-attention in the domain of the structured task sequence. SHAP analysis showed that the contribution of PFT features was about 42% of the total predictive significance, with the next contribution of Category (32%) and Semantic (26%) tasks. In order

to test generalization, the trained structure was also checked by external validation using the Pitt corpus of TalkBank (binary: Control vs Dementia). RF and Transformer models performed best on this independent dataset with a high accuracy of 91.4 and the highest ROC-AUC of more than 0.95, whereas SVM gave an accuracy of 90.1. The deep learning models (LSTM and GRU) showed a strong performance with ROC-AUC greater than 0.94, proving that multimodal feature integration is applicable to all datasets and diagnostic conditions. In general, the findings indicate that by combining acoustic and linguistic factors together with explainable AI, it is possible to achieve accurate and clinically interpretable and generalizable dementia detection. The suggested framework builds on the speech-based digital biomarkers by filling the gap between the performance optimization, interpretability, and clinical applicability in the real world.

Contents

List of Figures

List of Tables

Abbreviations

AD	Alzheimer's Disease
CTD	Category Fluency Test
EEG	Electroencephalography
EHR	Electronic Health Records
GRU	Gated Recurrent Unit
HC	Healthy Control
MCI	Mild Cognitive Impairment
PFT	Phonemic Fluency Test
ROC-AUC	Receiver Operating Characteristic – Area Under Curve
SHAP	SHapley Additive Evplanations
SFT	Semantic Fluency Test
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

In this chapter, the research problem that is presented is the speech-based dementia detection and the clinical and computational motivations that underlie this study. It describes the increasing world problem of dementia and the necessity of non-invasive solutions to early diagnostics. The chapter points out some of the gaps in the current AI-based systems, such as the absence of multimodal integration, interpretability, and cross-dataset validation. The question and objectives of the research are clearly stated, and the description of the proposed multimodal and explainable AI framework is given. Lastly, the chapter gives the structural organization of the thesis.

Dementia is one of the biggest worldwide public health issue, out of all of the diagnosed cases, approximately 60/70% are Alzheimer disease (AD) [1]. AD being a progressive neurodegenerative disease, it progressively deteriorates memory, executive functions, attention and language to cause severe cognitive and communicative loss. The consequences of late diagnosis are immense not just to the patients, but also to the caregivers, healthcare systems and the economy of countries. With the rising life expectancy of the global population, it could be estimated that the cases of dementia will grow significantly, which will only cause greater pressure on the availability and scalability of accessible screening options. Early diagnosis is still a clinical issue since early intervention can prevent cases of disease progression, maximize the therapeutic plan, and even enhance the overall quality of life.

Recent studies have enhanced this domain by incorporating explainable AI methods in detection of dementia. Shapley Additive Explanations (SHAP) as proposed by Oiza-Zapata and Gallardo-Antolina [2] to conduct feature selection and interpretability in speech-based AD detection enables the role of transparency in clinical decision-making. Recent studies have enhanced this domain by incorporating explainable AI methods in detection of dementia. Nonetheless, the traditional methods of diagnostics, such as neuroimaging, cerebrospinal fluid biomarker, and a whole neuropsychological testing are usually expensive, invasive, and unavailable in resource-limited locations. These constraints have inspired more research interest in speech and language analysis as a non-invasive, cost-effective, and scalable digital biomarker of early cognitive screening.

Expanding on this, Mekulu et al. [3] have proven the use of large language models (LLMs) that manage to capture subtle language variations in their narrative speech and use them to identify early-stage dementia. Jasodanand et al. [4] offered a comparable AI-based multimodal biomarker evaluation framework as well, demonstrating the possibility of using speech in conjunction with other clinical modalities. Additional references are Garcia-Gutiérrez et al. [5] who used machine learning-based speech analysis along the AD spectrum, and Jahan et al. [6] who concentrated on speech-based early dementia diagnosis with classical machine learning methods. Ding et al. [7] have performed a systematic review of AI methods, data, and issues in the field and emphasized the importance of interpretability and clinical testing.

Speech production is an activity that requires a high level of thinking as it requires the retrieval of memories, arranging them in a semantic order, accessing lexical information, planning activities at an executive level and motor control.The work by Vrindha et al. [8] and Qi et al. [9] reviewed the literature on the usefulness of spontaneous speech and linguistic markers as predictors of cognitive impairment by the researchers. Therefore, there are minor shifts in speech patterns, which often appear early in the process of cognitive impairment long before clinical symptoms become evident. Machine learning and deep learning techniques have shown promising results of discovering acoustic and linguistic predictors of

dementia in the past decade. Tang et al. [10] used linguistic characteristics based on automatic speech recognition, whereas Javeed et al. [11] performed a systematic review of machine learning-based methods of detecting dementia. Initial research was mostly based on manual creation of feature engineering and traditional machine learning models that are trained on textual and acoustic representations of surfaces. Though these methods demonstrated moderate success, they were based on the hand-selected features, which restricted the ability to generalize them to datasets and languages.

Moreover, there are a number of works on explainable linguistic and acoustic markers. Parsapoor [12] reviewed AI-based speech and language testing in dementia, and Shah et al. [13] reviewed language-agnostic speech representations based on domain knowledge to detect Alzheimer. In their works, Robin et al. [14] discussed the topic of progressive speech changes in patients with AD and Petti et al. [15] focused on ethical issues related to the implementation of AI to detect early signs of dementia. Lastly, Yang et al. [16] published a review of speech analysis in AD detection based on deep learning giving access to the state-of-the-art methods.

Although these significant progress has been made, there are still a number of methodological and translational gaps [16]. First, several of the reported systems have shown good performance with a single dataset but do not have serious external validation. The performance of models trained using a single corpus often drops when they are put in new demographic populations, different recording conditions, or language variation. This dependence on data may be viewed as a problem with reference to real-world generalization and scope clinical trust. Second, deep learning models including recurrent neural networks and Transformers are better predictors, although typically lack neuropsychological interpretability. Unless the reasoning behind AI-assisted diagnostic systems could be effectively aligned to established cognitive markers in a transparent manner, the clinicians would not have a desire to adopt these systems. Third, assessment measures also tend to be limited to general accuracy without reference to clinically significant levels of sensitivity and specificity needed to test kaggle dementia and mild cognitive impairment (MCI) screening.

This methodology designed by a two-dataset experimental design, which is structured, to overcome these limitations and hence to estimate both internal performance and cross-dataset generalization. The Kaggle dataset used is the Dementia Detection using Speech in which it serves as the main development corpus. This data includes multi-class labels (Healthy Control (HC), Mild Cognitive Impairment (MCI) and Dementia) and makes cognitive progression fine-grained modelling possible [17]. The multi-class design is especially significant since MCI is an intermediate stage that cannot be clinically easily identified, but which is very important to early intervention approaches. This dataset is used to train, tune and internally validate the model to learn discriminative acoustic linguistic representations in consecutive cognitive levels.

The Pitt Corpus of DementiaBank (TalkBank) is exclusively employed in external validation in order to determine robustness outside the development domain. Pitt data set has binary categories (Control vs Dementia) using standardized picture description procedures [18]. Notably, such data is not utilized in the course of training, eliminating the possibility of data leakage and maintaining an unbiased appraisal. The study explicitly quantifies domain shift resilience and representation stability by training models on an independent corpus with different demographic and different recording properties. Instead of stating single in-sample accuracy, this framework takes a look at the ability of learned multimodal features to maintain discriminative capacity in the face of real-world variability. This two-dataset approach will increase the rigor of the method and translational validity of the suggested system [19].

In this context, the study presents a multimodal speech representation, which is created based on standardized verbal fluency measures, such as Semantic Fluency Tests (SFT), Category Description Tests (CTD), and Phonemic Fluency Tests (PFT). The study does not engage the acoustic and linguistic modalities separately but rather builds a unified representation of 300 dimensions of features that retain task-relevant cognitive features. Acoustic characteristics encode temporal fluctuations, pause, pitch variation, articulation consistency and linguistic characteristics encode lexical abundance, semantic grouping, syntactic elaboration and

word retrieval patterns [20]. The model uses the synergistic data indicative of lower-order speech production and upper-order cognitive organization by combining these modalities.

In order to learn sequential dependencies between verbal tasks, recurrent neural networks like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are used. These architectures can represent inter-task temporal dynamics, which is an expression of neuropsychological fact that cognitive decline is manifested progressively instead of on discrete markers [21]. Simultaneously, Parallel self-attention models using Transformers are explored to evaluate the character of the world contextual modeling to improve the sensitivity to finer linguistic decay. The classical methods of machine learning, such as Random Forest and Support Vector Machines are also assessed to serve as a controlled comparative baseline in a single experimental system. The presented comparative analysis allows systematic evaluation of architectural trade-offs when preprocessing and feature engineering pipelines are the same.

The modeling architecture is directly modeled to support the interpretability of its components based on explanation mechanisms of SHAP. Local (instances-level) and global (data-set level) interpretations are produced that measure the contribution of features and conform predictive behavior with well-known neuropsychological constructs [22, 23]. This unification does not only go beyond the post-hoc visualization but also creates a clear connection between machine learning prediction and clinically meaningful speech markers.

In order to meet robustness, interpretability, and clinically calibrated performance, a deployable dementia screening tool needs to meet all three criteria at the same time. Thus, not only accuracy is a criterion of evaluation of models in this study, but also sensitivity, specificity, F1-score, and ROC-AUC, and, most importantly, sensitivity thresholds suitable during early cognitive screening. Together, this study is a step in the right direction toward building a reliable, scalable, and clinically consistent AI system to detect dementia through speech. The study covers algorithmic and translational aspects of AI in healthcare using multimodal

acoustical, linguistic, fusion, sequential deep learning models, Transformer-based contextual models, SHAP-based interpretability, and cross-dataset validation.

1.1 Background

Dementia, and especially the Alzheimer disease (AD) is one of the rapidly developing healthcare crisis in the world, which is experienced by a progressive communicative and cognitive loss. With the life expectancy on the increase, dementia prevalence keeps rising at a huge burden to the patients, caregivers, and healthcare systems. Clinical necessity of early detection is based on the fact that early intervention can slow down, enhance care planning and increase the quality of life of patients. Traditional methods of diagnosis however which include neuroimaging, cerebral biomarkers, and large neuropsychological batteries can tend to be expensive, invasive, and not available in low-resource settings. By comparison, speech production involves multifaceted cognition as it encompasses retrieval of semantic memory, executive planning, lexical access and time coordination. Early cognitive decline can be characterized by subtle changes in the acoustic stability, the patterns of pause, lexical diversity, and syntactic organization. These attributes make speech and language analysis an exciting non-invasive non-digital biomarker of scalable dementia screening.

The integration of XAI as shown in Fig 1.1 with multimodal feature fusion in this study is a deliberate and principled design choice motivated by two complementary clinical requirements. Feature fusion alone addresses representational completeness — since dementia simultaneously degrades motor speech control (acoustic domain), language production (linguistic domain), and executive timing (temporal-prosodic domain), no single modality captures the full diagnostic picture. However, fusing multiple modalities without interpretability creates a high-dimensional opaque decision space that clinicians cannot audit or trust. XAI via SHAP resolves this by decomposing the fused feature contributions into individual, ranked attributions that correspond to clinically recognisable neuropsychological

markers. Together, multimodal feature fusion maximises predictive completeness while SHAP-based XAI restores the transparency necessary for clinical adoption — neither component alone fulfils both requirements simultaneously.

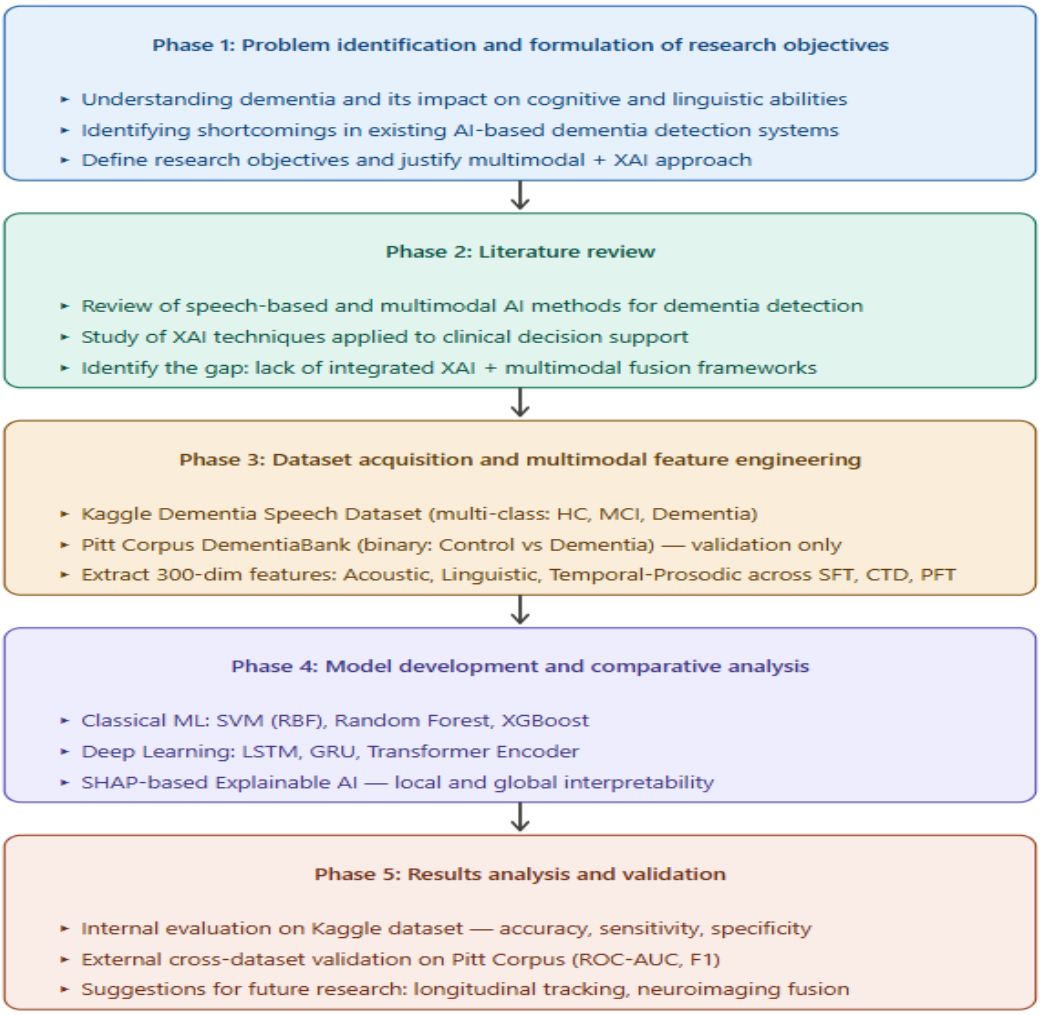


FIGURE 1.1: Research Methodology Phases

1.2 Problem Statement

Although the advances in artificial intelligence use to detect dementia are rapid, the issues have some major limitations that are still not addressed.

To start with, most of the current systems are more about identifying dementia in late stages, whereas identifying dementia at its early stage and modeling progression is poorly covered. MCI as a transitional stage between normal aging and

dementia is also one of the forms that are quite hard to detect because of the lack of pronounced and homogenous changes in speech. Lack of proper modeling of this phase constrains the clinical utility of AI-based screening devices. Second, many of the previous researches are based on either acoustic characteristics (e.g., pitch variation, speech rate, jitter) or linguistic characteristics (e.g., lexical richness, syntactic complexity). This unimodal method might not give much consideration to complementary information since cognitive impairment is found in various dimensions of speech production. Lack of built in multimodal modeling is a limitation to the richness of representation.

Third, most of the effective deep learning models are black-box models. They can be as precise as competitive but offer little transparency into what speech characteristics they use to make predictions. In clinical settings, interpretability leads to lack of trust, impedes adoption, and integration into decision-support workflows.

Lastly, interdataset generalization is still a significant issue. The use of models that have been trained on a specific corpus often does not work well when tested on new data because of demographic differences, recording conditions, linguistic variability, and task heterogeneity. Reported performance can be highly exaggerated with no cross-dataset validation to address the actual applicability in the real world.

All these restrictions lead to the desire to have a powerful, multimodal, interpretable, and generalizable AI system of detecting a speech-based dementia.

1.3 Research Questions

This study seeks to answer the following research questions in order to fill in the identified gaps:

RQ1: To what extent does the integration of acoustic, linguistic, and temporal-prosodic speech features into a unified multimodal representation improve the

discriminative performance and clinical interpretability of AI-based dementia detection, compared to conventional unimodal approaches?

RQ2: How effectively can a multimodal speech-based AI framework, augmented with SHAP-based explainability, distinguish between cognitively healthy individuals and those with dementia, and what are the diagnostic boundaries when applied to the transitional Mild Cognitive Impairment (MCI) stage?

RQ3: To what degree do the multimodal speech-derived representations and SHAP-based explanations generalize across heterogeneous datasets and elicitation paradigms, and do the resulting feature attributions align with established neuropsychological theory to support clinically trustworthy decision-making?

1.4 Research Objective

The study aims to answer the following objectives in order to answer these research questions.

i. The first aims at establishing a multimodal AI-based system, which can recognize dementia through speech and language characteristics based on standardized verbal fluency assessments. The design of the system is aimed at backing up acoustic, linguistic and temporal features into a single form of representation that maintains task specific cognitive signatures.

ii. The second goal is to differentiate stages of cognitive impairment progression, that is, to draw the line between Healthy Control, Mild Cognitive Impairment, and Dementia in a multi-class experimental environment. This performance feature deals with the very important issue of early detection.

iii. The third goal is to explicitly combine acoustic and linguistic characteristics in order to examine their interaction to predictive power and interpretability. The research does not consider modalities individually, but rather it builds up a consolidated feature space to represent complementary information.

iv. The fourth goal is to involve explainable artificial intelligence approaches, namely SHAP-based local and global interpretability, into connecting model predictions to clinically significant speech indicators. This will provide transparency and facilitate trust in clinicians.

v. The fifth aim is to prove model generalizability on various datasets. The experiment is run on the Kaggle multi-classes dataset, the internal development and staging is done, and external validation using the independent Pitt corpus (binary classification). This cross-domain robustness can be evaluated rigorously in this dual-dataset design.

1.5 Scope

In order to operationalize these objectives, the study is designed to follow a structural experimental design. The Dementia Detection using Speech Corpus is used [17], which allows one to use the data as multi-class for predicting the cognitive progression. The dataset can be used in training, hyper parameter optimization, and internal evaluation of Healthy and MCI and Dementia classes.

The Pitt Corpus of DementiaBank [18] is utilized solely to validate the external validity in order to measure the generalization and eliminate the bias related to the dataset. The binary labels (Control vs Dementia) of this corpus are based on the standardized narrative tasks. The study isolates training and validation data sets to make sure that the performance indicators describe transferable cognitive-linguistic representations rather than memorizing data in the datasets.

With this dual-stage assessment plan, the framework deals with a classification performance at the internal classification level and robustness at the external level, respectively, which makes the framework more translational.

Overall, this chapter laid the background of the research, as it defined the research problem, identified important gaps in the current dementia detection systems, and

developed research questions and objectives. It was well explained why acoustic and linguistic elements were combined with explainable AI. Moreover, the importance of cross-dataset validation in order to guarantee generalizability was pointed out. Chapter two is a literature review whereby the current literature is examined to place this work in the wider scope of science.

Given the multi-class complexity of the Kaggle dataset and the binary nature of the Pitt corpus, the study anticipates stronger dementia-control discrimination than fine-grained MCI staging, acknowledging MCI as a transitional diagnostic challenge rather than a fully solvable classification problem within the current framework.

This scoping decision ensures that reported findings are interpreted within realistic clinical constraints rather than inflated performance expectations.

1.6 Contributions

This study makes several significant contributions to the field of AI-based dementia detection.

i. Proposes a unified multimodal framework that integrates both acoustic and linguistic speech features to improve diagnostic accuracy and robustness.

ii. Conducts a structured comparison between classical machine learning models and deep learning architectures within a consistent experimental pipeline.

iii. Incorporates XAI techniques to enhance interpretability, increasing clinical trust and transparency of model decisions.

iv. Validates model generalizability through a dual-dataset strategy, including internal evaluation on the Kaggle dataset and external validation on the Pitt dataset.

v. Advances the development of clinically relevant, interpretable, and generalizable speech-based dementia detection systems.

1.7 Thesis Structure

The rest of this thesis is organized into six structured chapters. Chapter 1 introduces the research background, problem statement, objectives, and overall contributions of the study in the context of AI-based dementia detection. Chapter 2 presents a comprehensive literature review covering speech-based biomarkers, multimodal learning approaches, and explainable AI techniques. Chapter 3 details the proposed methodology, including dataset description (Kaggle and Pitt), preprocessing, feature extraction, fusion strategy, and model architectures. Chapter 4 describes the experimental setup and presents the results obtained from both internal and external validation. Chapter 5 provides an in-depth discussion, theoretical interpretation, limitations, and clinical implications of the findings. Finally, Chapter 6 concludes the thesis with key contributions, and directions for future research.

Chapter 2

Literature Review

This chapter summarizes the currently existing research in AI-based dementia detection, speech biomarkers, multimodal feature integration as well as explainable artificial intelligence. It reviews classical methods of machine learning, deep learning models, and the recent breakthroughs in the field of the large language models used to analyze cognitive impairment. Special interest is put on the studies that deploy acoustic and linguistic features in the early detection as shown in Fig 2.1. The interpretability frameworks are also reviewed in this chapter and some difficulties associated with dataset variability and clinical implementation are discussed.

2.1 Evolution of Speech-Based Baselines and the Generalizability Challenge

Precursor studies in the present window have framed classical and deep learning on top of engineered acoustic features as a viable speech-based dementia detection baseline. Extensive testing in both machine learning and deep architectures pointed to the fact that robust pre-processing, thoughtful feature selection and balanced evaluation procedures tend to be much more important than pursuing complex models, particularly when dataset sizes are small or heterogeneous [19].

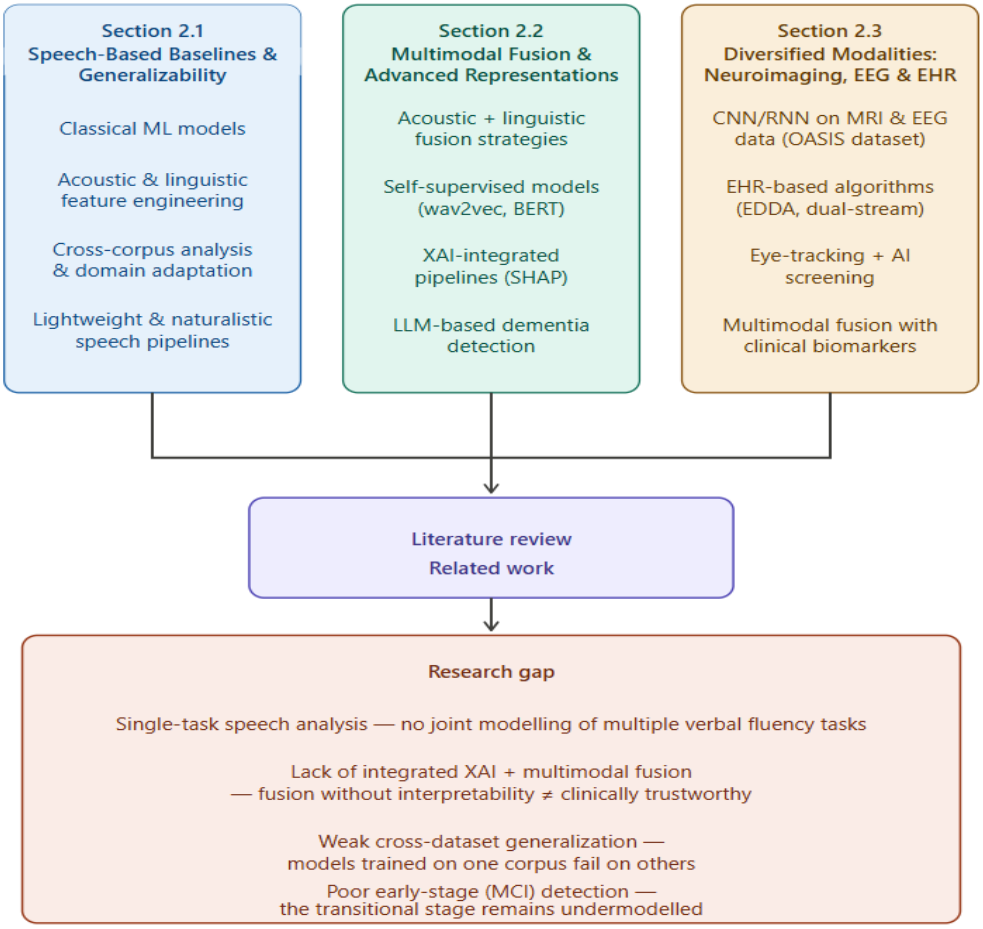


FIGURE 2.1: Outline of Literature Review

Based on this evidence base, a systematic review mapped the speech-and NLP-based pipelines end-to-end, that is, tasks, features, models, and validation pitfalls, and urged a more rigorous comparability across studies and explicit reporting of clinical relevance [20].

The major technical issue that appeared was that of generalizability. Cross-corpus analysis revealed that models which are trained on one dataset often fail at another due to the incompatibility of language, task, and recording conditions, which highlights the importance of domain adaptation and protocol harmonisation [21].

Simultaneously, a research showed that, with sensitive acousticlinguistic engineering and appropriate model choice, it is possible to achieve competitive performance on naturalistic speech, inspiring pipelines that are both lightweight and useful in clinical practice as provided in the following paper [22]. These messages were

supported by a wider scoping review of AI in early dementia detection that suggested prospective validation, explainability, and alignment with clinical pathways to enhance the bridge between the bench and bedside gap in provided article [23].

Having clinical translation in mind researchers experimented with tool chains that can be deployed in practice. One of the streams assessed AI tools quantifying language and speech patterns in the case of Alzheimer disease, noting opportunities but also warning of biases, small cohorts, and inadequate reporting on fairness and interpretability [24].

A second seminal direction that used large language models as a substitute to spontaneous speech/text demonstrated that pretrained language representations could reveal dementia related signals without intensive feature engineering, whilst posing questions about transparency and data demands for the dimentia [25].

In complementary fashion, a distance smart phone screening system showed that scalable, low-burden assessment outside of the clinic was possible, a milestone that is significant in terms of real-world implementation [26]. Explainability went hand in hand with: transformer-based transcript models with XAI methods demonstrated how token-level attributions could lead to better faithfulness of automated judgments by highlighting clinically meaningful lingo features [27].

Basic resources and standards jump-started improvements. An extensively deployed spontaneous speech corpus and challenge data framed common tasks, task splits and task metrics, allowing the comparison of apples-to-apples, and the iteration of acoustic and linguistic baselines quickly [28]. The modeling side, bottleneck neural networks and aggressive data augmentation were found to enhance robustness to recording variability, prioritizing representation learning that learns disease-salient cues and regularizes nuisance variability [29].

Classic sets of paralinguistic features (e.g., MFCCs/log-mel) were still robust workhorses; systematic evaluations affirmed their usefulness and explained trade-offs between spectral, prosodic and voice-quality features in low-resource environments [30].

Fusion of modalities within speech pipelines further improved accuracy. Experiments on temporal integration of transcripts with acoustics revealed that complementary cues, what is said and how it is said, enhance classification when co-focused over discourse structure and time indicating sequence models that are sensitive to the dialogical dynamics [31].

In addition to classification accuracy, semantic discourse tasks also revealed subtle lexical-semantic degradation in cognitively impaired populations, which encourages the addition of narrative/topic-level probes to protocols [32]. Studies of linguistic features, particularly lexical richness, syntactic complexity and measures of coherence showed that well-maintained feature banks with the assistance of traditional classifiers could also compete with deep models when presented with limited amounts of data [33].

Big data on the real world reinforced clinical validity. Another interesting cohort study, which is based on longitudinal voice recordings, stated that end-to-end deep learning can identify dementia-related patterns in natural speech, which highlights the potential of passive or opportunistic monitoring at scale [34].

In under-resourced scenarios, low-resource scenarios led to clarified data-efficient approaches (self-supervision, transfer learning, multilingual pretraining), which are crucial to ensure fair cross-language and cross-dialect deployment [35]. The hybrid approaches, which combined self-supervised acoustic encoders (e.g., wav2vec) with contextual language models (e.g., BERT), embraced not only non-semantic voice signatures but also semantic impairments, which has served as a strong template in modern multimodal-within-speech architectures [36].

Combined acoustic-and-language deep learning pipelines repeatedly outperformed unimodal systems, especially when trained with careful regularization and tested on subject-wise splits avoiding leakage [37]. Beyond speech, multimodal models that combine clinical, neurocognitive or sensor-based information were used to demonstrate how explainable, layered models can be used to aid detection and, in fact, prediction, aligning the outputs of an algorithm with concepts that are comprehensible by a clinician [38].

Lastly, initial experiments on automated screening under free and natural conditions, e.g. telephone-quality or in the field recordings indicated the possibility of inexpensive front-line tools, although it demanded strenuous QA, calibration and human in the loop validation prior to clinical application [39].

2.2 Multimodal Fusion and the Role of Advanced Representations

Recently developed AI-based dementia detection is highly motivational towards learning the explainable, translationally ready multimodal frameworks. Shao et al. presented a model of FFG that was tested on three publicly available assessment datasets, namely Pitt, ADReSS, and ADReSSo, and exhibited better results on all the benchmarks and achieved 85.85% and 84.30% accuracy on Pitt and ADReSSo datasets, respectively [40].

To supplement the speech based modalities, Abdulaal et al. proposed a decoupled frequency-spatial attention model of EEG analysis that combines the continuous wavelet transform and artifact subspace reconstruction to extract optimally features out of 19-channel recordings. Their approach obtained 96.7% accuracy on AD vs. healthy controls, 92.3% on FTD vs. healthy controls and 87.6% on AD vs. FTD classification on the Miltiadous dataset, which is better than traditional machine learning and current deep learning models [41].

In addition to signal-specific architectures, machine learning-based early detection methods have also been widely studied on neuroimaging data, e.g., OASIS. Ensemble and classical models such as SVM, Logistic Regression, AdaBoost and MLP have also been optimized to have the highest levels of diagnostic accuracy, sensitivity, and specificity regarding dementia classification [42]. On a similar note, the comparative analysis of eight state of the art AI/ML models on MRI based OASIS data reported SVM as the best-performing model with an accuracy of 92, a specificity of 0.92 as well as an F1-score of 0.91 which MMSE and CDR identified

as the strongest predictive features using LASSO and feature selection by Fisher exact test [43]. The speech-based systems like CogniWave also note the diagnostic potential of the acoustic biomarkers with a 92% accuracy, 96% precision, and 90% recall with cross-validation strategies. [44]

Despite these advances, a critical gap persists in the existing literature: studies that employ multimodal feature fusion rarely integrate systematic explainability mechanisms, and conversely, XAI-focused studies often operate on single-modality features. This bifurcation limits translational utility fusion without interpretability remains clinically untrustworthy, while XAI without multimodal input provides incomplete diagnostic coverage.

The present study is specifically designed to bridge this gap by co-designing the feature fusion and XAI pipeline as an integrated system rather than treating interpretability as a post-hoc addition.".

2.3 Diversified Modalities: From Neuroimaging and EEG to EHR Integration

AI has been implemented in clinical and real-world health care environments in opportunistic and large-scale screening. The Emergency Department Dementia Algorithm (EDDA) is an EHR-based algorithm that was created and tested to detect risk of dementia in older adults in emergency departments [45], and the dual-stream algorithms that use structured and unstructured data in EHR have been suggested to estimate prevalence and detect dementia more effectively [46].

Multimodal data based hybrid deep learning systems have been shown to be more predictive on early detection of Alzheimer [47], and extensive surveys indicate that CNNs- and RNNs-based networks can obtain up to 96.0% predictive accuracy on Alzheimer and 84.2% on mild cognitive impairment [48]. New modalities including eye-tracking alongside AI also yield positive outcomes with a 88% accuracy, 85% sensitivity and 86% specificity [49].

Wider surveys highlight the disruptive nature of AI, with vendors being ML, DL, and NLP, in the diagnosis and treatment of neurological conditions, such as Alzheimer's disease, stroke, brain tumors, and also note integration issues and ethical concerns in clinical implementation [50]. Deep learning systems based on the use of EEG have shown strong accuracy in diagnosing Alzheimer with a high accuracy rate of up to 97.27 [51], but comparative research studies with cognitive screening tools (MMSE, RUDAS, SAGE, ADAS, MoCA) have shown the comparative diagnostic reliability of these tests in detecting the disease at an early stage [52]. More sophisticated multimodal neuroimaging fusion models like the YOLOv11 based MRI-dti fusion have reached 93.6% accuracy and 91.6% recall in distinguishing early Alzheimer [53].

Nonetheless, the size of EHR-based studies indicates the possibility of detection bias in dementia studies, and therefore, high evaluation protocols are necessary [54]. In-depth literature reviews of deep learning methods also indicate that there continue to be challenges regarding generalizability, imbalance in datasets, and interpretability [55].

Biomarker-based methods, such as plasma-based indicators, are promising in determining Alzheimer's proteinopathy before dementia occurs [56], whereas emergency department research indicates that cognitive impairment is under-recognized because of low screening and referral systems [57]. CNN networks that use EEG spectrograms have been further optimized to achieve cross-subject validation, achieving 79.45% (AD/CN) and 80.69% (AD +FTD/CN) validation [58].

Primary care has been established as a crucial environment toward the early identification of dementia especially with the aid of prevention measures and the use of standardized diagnostic measures [59]. The systematic reviews contain research over the last ten years indicating the increasing efficiency of multimodal and AI-driven fusion methods to improve the accuracy of the diagnostic process [60]. Classical screening tools like Brief Cognitive Scale prove to have discriminatory power between normal screening, MCI, and dementia although more clinical enhancements are needed [61]. EEG-based methods are also appealing because of their

affordability and non-invasiveness, and current studies cover the standardization of the data sets and the robustness of the models [62]. Lastly, reviews of machine and deep learning-based Alzheimer detection summarize progress in modalities, risk factor modeling, as well as clinical translatability metrics, supporting the necessity of explainable, multimodal, and clinically translatability frameworks [63].

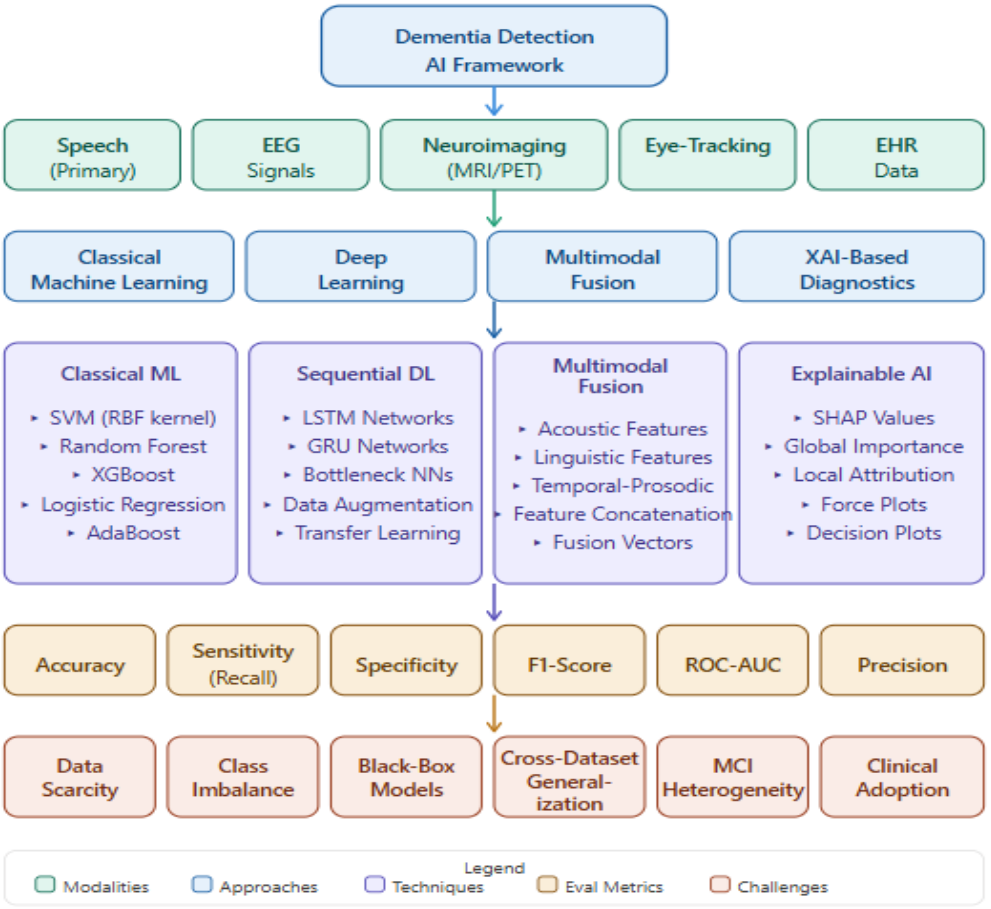


FIGURE 2.2: Taxonomy of AI-Based Dementia Detection Systems

Although there has been significant advances as shown in Fig 2.2. in the field of speech-based detection of dementia, the available research tends to be based on single-task analysis, in-interpretable, or it does not consider the clinical utility options. Further, minimal efforts have been done to model several verbal fluency tasks jointly and maintain cognitive relevance of these tasks.

The present research is attempting to fill these gaps by suggesting a task-sensitive, interpretable multimodal speech model that incorporates acoustic, linguistic, and

temporal cues as shown in table 2.1 across standardized fluency tests as well as explainable AI to support clinically-based decisions.

TABLE 2.1: Comprehensive overview of prior studies on AI-based dementia detection using speech and language analysis

Ref. No.	**Year**	**Methodology**	**Dataset Used**	**Limitations**
[2] Oiza-Zapata & Gallardo-Antolín	2025	SHAP-based feature selection with ML classifiers	Pitt, ADReSS	Small dataset, limited generalization
[3] Mekulu et al.	2025	Large Language Models (LLMs) for narrative speech	Narrative speech transcripts	High computational cost; bias
[4] Jasodanand et al.	2025	AI-driven multimodal data fusion (speech, imaging, biomarkers)	Alzheimer's cohorts	Data integration complexity
[5] García-Gutiérrez et al.	2024	Acoustic ML-based speech analysis	AD spectrum datasets	Demographic bias
[6] Jahan et al.	2024	Speech features with ML classifiers	Small clinical dataset	Small dataset, low external validity
[7] Ding et al.	2024	Survey of AI techniques, datasets, and challenges	ADReSS, Pitt, etc.	No experimental validation
[8] Vrindha et al.	2023	Review of speech and text-based AD detection	Multiple corpora	Review only

TABLE 2.1: Continued from the previous page

Ref. No.	Year	Methodology	Dataset Used	Limitations
[10] Tang et al.	2023	Explainable AI using linguistic features from ASR	ASR-transcribed speech	Dependent on ASR accuracy
[9] Qi et al.	2023	Review of non-invasive speech-based AD detection	Literature	Lacks empirical validation
[11] Javeed et al.	2023	Systematic review of ML for dementia prediction	Prior works survey	No experiments
[12] Parsapoor	2023	AI-based assessment of speech impairments	Clinical speech data	Limited interpretability
[13] Shah et al.	2023	Language-agnostic speech embeddings with domain knowledge	ICASSP challenge data	Limited languages
[14] Robin et al.	2023	Automated detection of progressive speech changes	Early AD speech	Limited to early stage
[15] Petti et al.	2023	Ethical analysis of AI for AD detection	N/A (conceptual)	No technical solution

TABLE 2.1: Continued from the previous page

Ref. No.	Year	Methodology	Dataset Used	Limitations
[16] Yang et al.	2022	Deep learning-based speech analysis review	Survey	Review only
[19] Kumar et al.	2022	ML and DL architectures for dementia detection	Small datasets	Overfitting; small size
[20] Ševčík & Rusko	2022	Systematic review using speech and NLP	Multiple datasets	Review only
[21] Ablimit et al.	2022	Cross-corpus dementia detection	ICASSP corpora	Dataset mismatch
[22] Bertini et al.	2022	Automatic classifier for spoken English	Spontaneous English speech	Language-specific bias
[23] Li et al.	2022	Scoping review on AI in dementia	Literature survey	No new experiments
[24] Favaro et al.	2022	AI-based language and speech pattern analysis	Clinical samples	Limited interpretability
[25] Agbavor & Liang	2022	LLM-based speech-to-dementia modeling	Spontaneous speech	Model bias
[26] Fristed et al.	2022	Remote AI system using smartphones	Remote patient speech	Device variability

TABLE 2.1: Continued from the previous page

Ref. No.	Year	Methodology	Dataset Used	Limitations
[27] Ilias & Askounis	2022	Transformer-based explainable dementia detection	Dementia transcripts	Black-box limitations
[28] Luz et al.	2021	Spontaneous speech recognition for AD	ADReSS	Limited size
[29] Liu et al.	2021	Neural networks with bottleneck features and augmentation	ICASSP	Needs more robustness
[30] Meghanani et al.	2021	Log-Mel spectrograms and MFCC features	SLT dataset	Narrow scope
[31] Martinc et al.	2021	Temporal integration of transcripts and acoustics	AD datasets	Limited multimodal scope
[32] Antonsson et al.	2021	Discourse task-based semantic analysis	Cognitive impairment speech	Task-specific only
[33] Calzà et al.	2021	Linguistic features with classifiers	Clinical samples	Feature selection bias
[34] Xue et al.	2021	Deep learning on voice recordings	Framingham Heart Study	Dataset demographic limits

TABLE 2.1: Continued from the previous page

Ref. No.	**Year**	**Methodology**	**Dataset Used**	**Limitations**
[35] Pappagari et al.	2021	Speech and language technology for low-resource settings	Interspeech challenge	Low-resource limits
[36] Zhu et al.	2021	Wav2vec and BERT (WavBERT) model	Interspeech dataset	High computation
[37] Mahajan & Baths	2021	Acoustic and language deep learning approaches	Spontaneous speech corpora	Limited validation
[38] El-Sappagh et al.	2021	Multimodal explainable AI model	Clinical multimodal data	High complexity
[39] Syed et al.	2020	Spontaneous speech ML classifier	Interspeech 2020	Limited dataset size

The literature review demonstrates the important progress made in the field of speech-based dementia detection and presents the most prominent shortcomings, specifically, multimodal integration, interpretability, and cross-dataset reliability. The models available though many have shown promising accuracy, few of them deal with early-stage detection in a holistic manner and have a clinically meaningful explanation. These loopholes explain the necessity of the proposed multimodal and explainable framework, which is described in the next chapter of the methodology.

Chapter 3

Research Methodology

The chapter entails the suggested methodological framework of multimodal speech-based dementia detection. It outlines such datasets used in the internal development and external validation, the Pitt corpus and the Kaggle multi-class dataset. It has a detailed explanation of the feature engineering process which includes acoustic, linguistic and temporal attributes.

The balancing of both classes strategy, preprocessing pipeline, balancing of classes, model normalization are systematically defined and the model architectures. Additionally, the XAI methods and the design of the experiment to perform cross-dataset validation are proposed.

The study was constructed as a multi-phase, multi-faceted computational system that would create an interpretable and practical (clinical) machine learning system to detect cognitive impairment through multimodal speech biomarkers. The data engineering, signal processing, statistical learning theory, deep neural network modeling and XAI are united in the methodological pipeline.

The goal was not only to enhance the level of predictive performance but to create a translucent, theory-consistent diagnostic aid which can be used to assist in making clinical judgments. The study is done in a systematic manner starting with raw speech records which are transformed into features, feature modeling, interpretability analysis and providing clinical validation. All the stages were well

planned shown in Fig 3.1 to make them reproducible, strong, and able to translate into real-life healthcare settings.

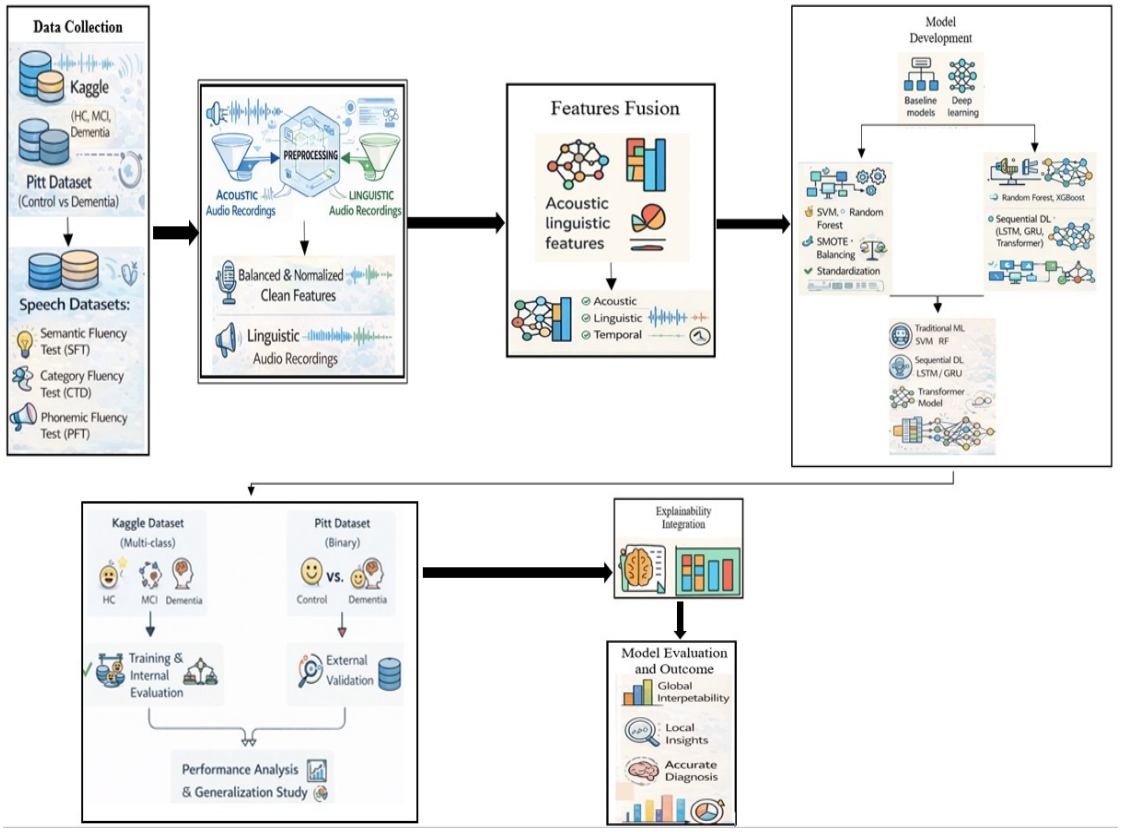


FIGURE 3.1: Stages of the proposed multimodal dementia detection framework

3.1 Multimodal Acquisition and Multimodal Feature Engineering of Data

This research is based on a curated dataset that was developed based on three verbal fluency standardized neuropsychological tests, namely SFT, CTD, and PFT. These tests find most applications in clinical neuropsychology to assess separate yet interconnected cognitive functions that are known to be impaired in neurodegenerative pathologies, including Alzheimer disease and other types of dementia.

This study combines all three tasks shown in Fig 3.2 a way that it only captures the complementary cognitive signatures as opposed to using a single behavioral measure.

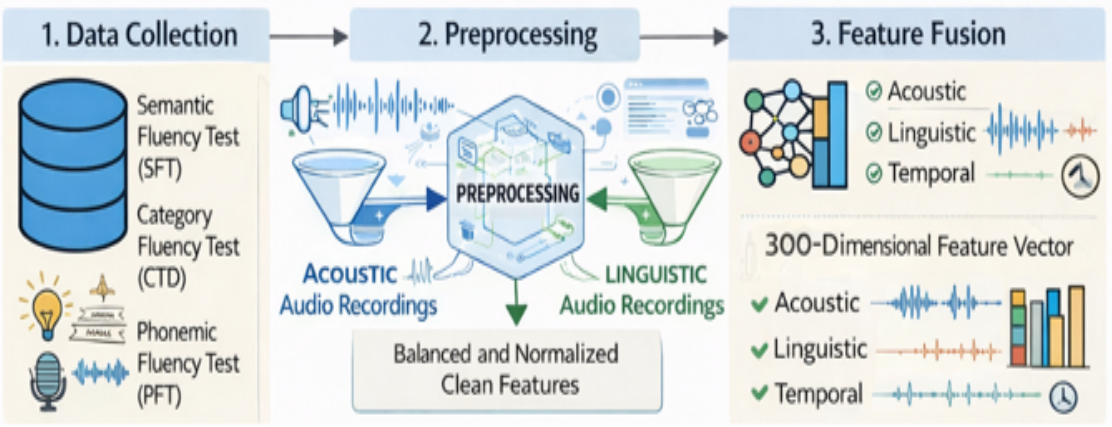


FIGURE 3.2: Overview of the three standardized verbal fluency tasks used for data collection

SFT involves the participants focusing on producing as many words as possible of a certain semantic category (ex. animals or fruits) within a time limit, which is normally sixty seconds. This exercise mainly tests semantic memory which can be described as organized knowledge on categories, concepts and connection between objects.

Semantic memory impairment is a characteristic of the early stages of Alzheimer disease, which is commonly observed in the form of decreased lexical variety, word repetition and lack of ability to group semantically related objects. To the neuroanatomy point of view, semantic fluency performance is linked closely to the integrity of the temporal lobe, especially in areas that deal with conceptual representation.

CFT also examines the executive control processes and cognitive flexibility. Though it is of a similar structure with the semantic task, it puts more emphasis on the switching strategies and the organized retrieval. Planing, inhibition, and switching of tasks Executive function are controlled to a large part by frontal lobe networks, and can only be adapted through them. In a person who is cognitively impaired, executive dysfunction comes out as a long pause, lack of efficient switching between subcategories and low retrieval speed. Hence, speech generated in the process of this activity contains important temporal and strategic data that can be used as a biomarker of early impaired cognitive functioning.

PFT involves the participants giving words of a given letter. In contrast to the semantic fluency, the phonemic fluency is highly determined by phonological retrieval strategies and inhibitory control. Frontal-subcortical circuits are involved in this task and are particularly vulnerable to the effects of executive dysfunction.

The semantic impairment is not as evident or earlier than the phonemic fluency deficits in many cases of neurodegenerative conditions. In turn, the inclusion of PFT allows one to determine subtle executive and phonological impairments that could be hidden when using semantic activities only.

Out of the recorded samples of the speech, a comprehensive feature engineering was performed. One hundred features were taken out of each task making a total of three hundred features per participant. These characteristics were categorized into three main modalities namely acoustic, linguistic and temporal-prosodic modalities. The acoustic parameters were calculated directly out of the speech wave form and they are basic frequency (pitch), jitter, shimmer, harmonic ratio to noise and spectral attributes. The measures are indicative of neuromotor control of speech production because vocal instability and micro-variations in frequency may be signs of neurological impairment in motor coordination.

Natural language processing was employed to derive the linguistic features. They contain measures of lexical diversity (i.e. type-token ratio, syntactic complexity measures, mean utterance length, and part-of-speech distributions).

Lexical diversity reflects native terms vocabulary richness and semantic accessibility, which reduce in the case of cognitive loss. Early dementia is also characterized by syntactic simplification and low grammatical variability.

Temporal-prosodic features are features that measure aspects of the timing of speech such as speech rate, articulation rate, pause period and pause frequency. The pause analysis is also very informative as more pauses are usually associated with difficulty in lexical retrieval or executive impairment. A combination of these three modalities makes the study have a complete representation of the speech, both neurocognitive and neuromotor processes.

The data of the subjects was coded as fixed order concatenated 300 dimensional feature vectors where features 1-100 are SFT, 101-200 CTD and 201-300 PFT.

The speech dataset was organized into three structured cognitive tasks: Semantic Fluency Test (SFT), Category Fluency Test (CTD), and Phonemic Fluency Test (PFT). Each task was represented using a fixed index range (SFT: 1–100, CTD: 101–200, and PFT: 201–300) to maintain consistent feature grouping and simplify model processing. This indexing strategy ensures clear separation of task-specific speech features while preserving their sequential representation within the dataset. Additionally, the structured segmentation allows the model to learn task-dependent linguistic and phonemic patterns that are known to be associated with cognitive decline.

Organizing the dataset in this manner improves feature management, enhances interpretability, and facilitates multimodal feature fusion across different cognitive speech tasks.This graphical representation as shown in Fig 3.3 does not alter the identity of tasks and can still be compatible with the machine learning algorithms.

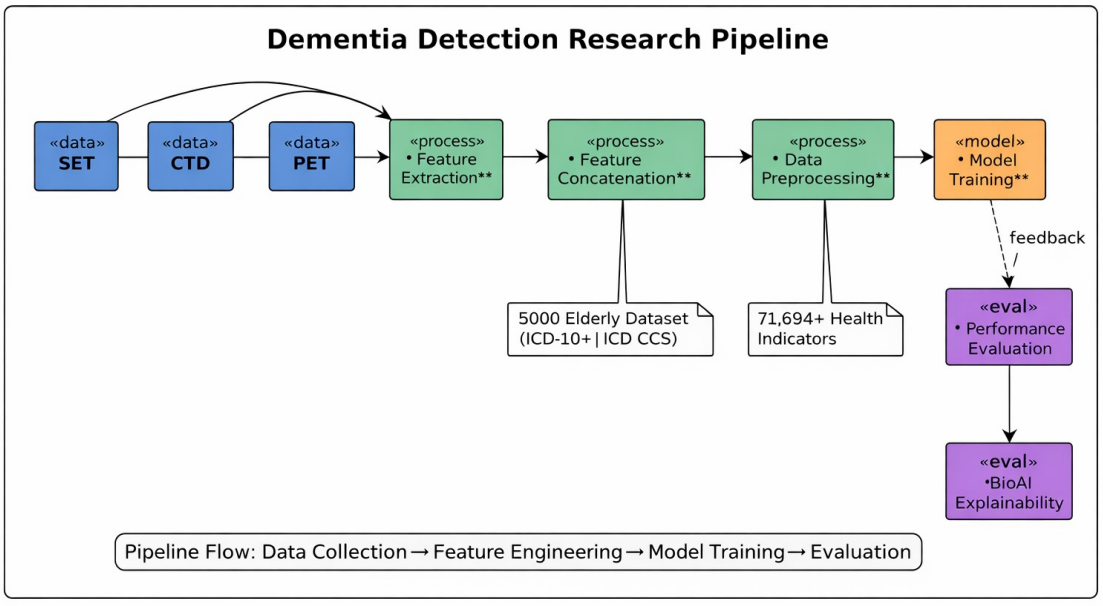


FIGURE 3.3: Multimodal feature engineering and vector construction process

Separate preprocessing pipelines were used on each dataset in isolation to prevent cross dataset contamination as well as maintain experiment integrity. In order to

validate the model, as well as to develop it, two separate speech sets were used. Datasets to ensure robust evaluation and generalization of the proposed dementia detection framework. The Kaggle dataset serves as the primary dataset and contains multi-class cognitive labels, including Healthy Control (HC), Mild Cognitive Impairment (MCI), and Dementia. This dataset is used for model training and internal evaluation as design shown in Fig 3.4. In contrast, the Pitt Corpus dataset is used as an external validation dataset and includes binary classification labels (Control vs Dementia). The use of two datasets enables the evaluation of model performance under different data distributions and recording conditions. This cross-dataset validation strategy strengthens the reliability of the proposed framework and ensures that the model does not overfit to a single dataset.

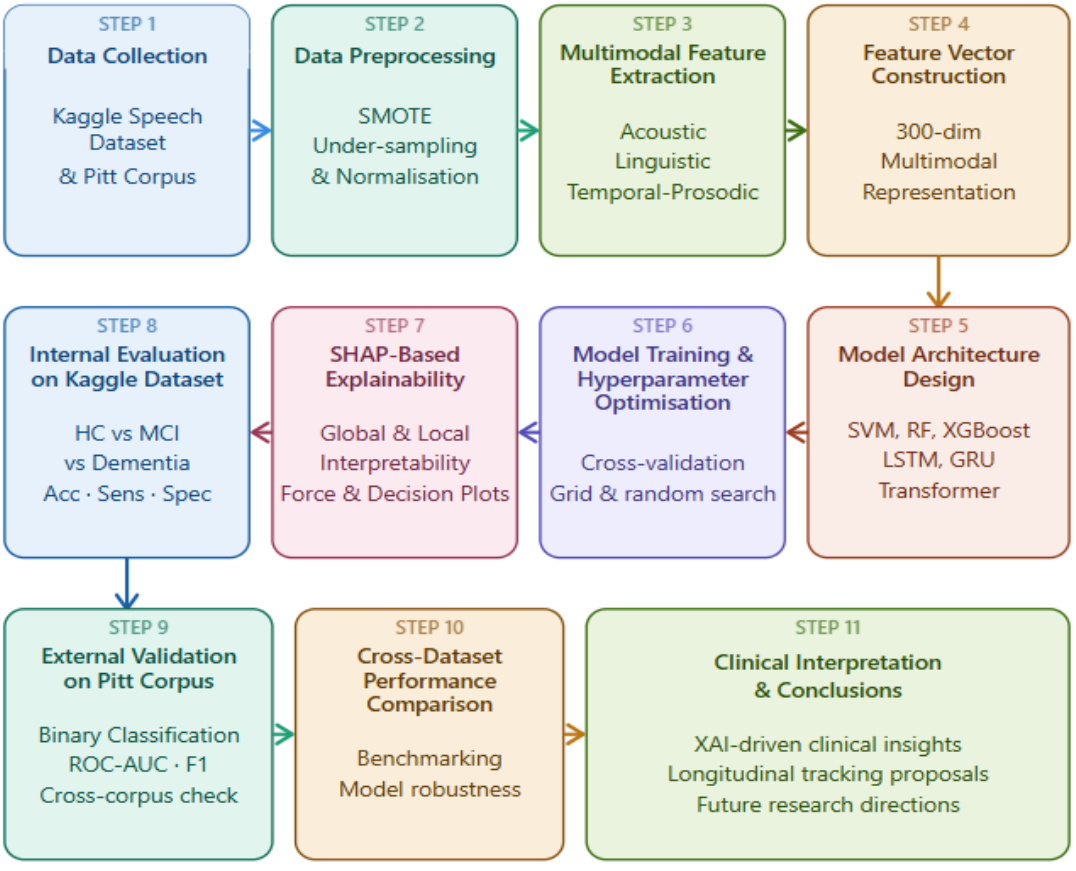


FIGURE 3.4: Steps of Research Design

In order to evaluate model robustness and cross-domain generalization, an external sample was used, the Pitt Corpus of DementiaBank (TalkBank). In contrast to the Kaggle one, the Pitt corpus has binary class labels (Control vs Dementia).

This data was exclusively reserved to do independent validation and was not used during the course of training to avoid data leakage and unbiased assessment.

The two datasets are used to fulfill two methodological purposes. To begin with, the multi-class Kaggle dataset can be used to test the fine-grained stages of cognitive decline, especially the difficult MCI category. Second, the Pitt corpus makes it possible to test whether the acquired acoustic and linguistic representations are generalized to the area of initial training. Such a design of two datasets enhances the scientific validity of the study by showing not only the intra-dataset performance but also the ability of the study to generalize to other datasets.

3.2 Data Preprocessing and Experimental Design

FIGURE 3.5: Preprocessing pipeline applied

The dataset was preprocessed as shown in Fig 3.5 extensively before Data preprocessing was performed to ensure the quality and reliability of the speech dataset before model training. Initially, missing values were examined and appropriately handled to maintain dataset consistency, and outlier detection techniques were applied to identify abnormal or extreme values arising from recording noise or feature extraction variations, preventing negative effects on model performance.

Subsequently, feature reduction was conducted to eliminate redundant or less informative attributes, reducing dimensionality and improving computational efficiency while retaining the most relevant information. Feature engineering was then carried out to extract meaningful acoustic and linguistic characteristics—including pitch, pause duration, speech rate, lexical richness, and fluency—which were further organized into phonemic, semantic, and category-based groups and integrated through multimodal feature fusion to comprehensively represent speech patterns associated with cognitive decline.

To address class imbalance, a mixed sampling strategy was adopted: SMOTE (Synthetic Minority Over-sampling Technique) generated synthetic samples for the MCI and Dementia classes, followed by random under-sampling of the Healthy Control class to ensure balanced representation across diagnostic categories. All features were normalized using StandardScaler to bring them to zero mean and unit variance, which is essential for models such as Support Vector Machines or neural networks that are sensitive to feature magnitudes.

Finally, the dataset was split into stratified 80/20 train-test subsets, preserving class proportions, and the model was trained with hyperparameter optimization, while the held-out test set provided an unbiased evaluation of generalization performance.

3.3 Architectural Implementation and Model Development

A comparative modeling framework has been embraced to assess the predictive performance comprehensively and covers both the classical machine learning algorithms and the deep learning architectures. By this means, it will be possible to evaluate whether sequential modelling or non-linear optimization of boundaries are more effective in multimodal speech data. The Support vector machines were trained with a Radial Basis Function kernel in order to identify the non-linearities in the high dimensional feature space.

SVM aims at determining an optimal hyperplane that values maximum margins between classes and reduce classification error. The data to a dense dimensional space to allow the division of intricate patterns which cannot be separated on a linear scale.

Random Forest classifiers were used as ensemble models which were made up of many decision trees that were trained on bootstrapped subsets of the data set. Independent predictions are made by each tree with the ultimate classification being achieved by majority voting. Random Forests are noise resistant, overfitting resistant, and have inherent feature importance values, which are useful in terms of interpretability.

XGBoost which is a gradient boosting framework was also used. In comparison to bagging-based ensembles, boosting approaches train trees sequentially, in a way that the subsequent model attempts to correct the residual errors of the previous model. Terms of regularization were added in order to avoid overfitting, and hyperparameters were optimized with the help of Bayesian search methods.

Recurrent neural networks were built to model explicitly the serial relationship between SFT, CTD and PFT tasks. The LSTM networks apply gated network to store the long-term dependencies and reduce the vanishing gradient problem.

GRU have a simpler architecture and few parameters but they are effective in terms of memory retention. This input was restructured into three time steps, which comprised of 100 features of a single task and as a result, the network learnt progression patterns among tasks.

Transformer encoder architecture was also tested. Transformers have self-attention models that calculate weighted interactions of all features at the same time. This facilitates the dependency modeling of the world without recourse to the sequential recurrence. Transformers however are generally data intensive, since their parameters are usually complex to obtain good performance.

Each model was set up with multi-class classification with categorical cross-entropy loss and early stopping to eliminate overfitting.

3.4 Explainable Artificial Intelligence Implementation

Considering the stakes of clinical decision support, the transparency of the models was given priority. The inclusion of SHAP-based Explainable AI in this framework is not merely a post-hoc analytical tool but a core design requirement arising from the multimodal integration itself. When acoustic, linguistic, and temporal-prosodic features from three separate verbal fluency tasks are fused into a single 300-dimensional representation, the resulting feature space becomes inherently complex and opaque. Without a principled interpretability mechanism, it is impossible to verify whether the model's predictions are driven by clinically meaningful speech patterns or by spurious statistical correlations in the training data.

SHAP resolves this ambiguity by providing mathematically grounded Shapley values that attribute each prediction to specific contributing features. This integration ensures that the diagnostic system is not only accurate but also auditable — a prerequisite for any AI tool intended for clinical decision support The SHapley Additive exPlanations model was incorporated in order to offer both local and global interpretability. The SHAP is based on cooperative game theory, which considers every feature as a player that helps to make the final prediction. Shapley value The Shapley value is a measure of the average marginal contribution of a feature to the potential combinations of features. The interpretation globally was obtained by calculating mean absolute SHAP values on the test set, which resulted in a ranked list of importance.

In general, this chapter outlined the systematic step-by-step process of data collection up to model development and integration of interpretability. There was a multimodal characteristic representation and dual-data validation plan which was meant to guarantee predictive performance and generalizability.

The framework integrates classical machine learning, sequential deep learning, and Transformer-based models with SHAP-based explanations, which forms a solid basis of examining the experiments, which are discussed in the following chapter.

Chapter 4

Experimental Results

This chapter will show the experimental outcomes of the internal and external analysis of the suggested framework. The rating of multi-class classification was evaluated using the Kaggle dataset in the category of Healthy Control, Mild Cognitive Impairment, and Dementia. The Pitt corpus was later used to externally validate the results to test the generalization in a binary classification condition.

Such performance measures as accuracy, sensitivity, specificity, ROC-AUC, and precision-recall analysis are reported. Systematic analysis of comparative analysis is conducted between machine learning and deep learning architectures.

4.1 Evaluation Strategy

The experimental evaluation process was to be conducted in order to strictly evaluate predictive performance, robustness, and clinical reliability of the proposed multimodal speech-based diagnostic framework. Each of the models was tested on a held-out test set produced by an 80/20 stratified split to provide a fair estimation of the performance in generalization. The evaluation protocol was particularly concerned with the general classification accuracy as well as clinically relevant measures of class-specific diagnostic measures as follows; sensitivity (recall), specificity, precision (positive predictive value) and F1-score.

Accuracy is not enough in medical decision support systems, especially in systems that deal with neuro-degenerative disorders, the cost of false negatives and false positives remains to be explicitly studied to find out how practical it is in the real world.

Since the classification task was multi-class, i.e., it included (HC), (MCI), and Dementia, the performance was considered on a global level and on a class-by-class basis. Each model was then used to create confusion matrices that could be used to visualize the misclassifications pattern and also to understand any systematic weaknesses in the diagnostic. The cross-validation in training further made sure that hyperparameter optimization did not overfit the test.

4.2 Performance Analysis of the Model Comparative

The relative analysis of the models that were adopted showed a definite order of prediction performance in Table 4.1.

TABLE 4.1: Comparative Performance of Models on Kaggle Multi-Class Test Set.

Model	**Recall**	**Precision**	**F1-Score**	**Accuracy**
GRU	0.50	0.778	0.609	0.762
SVM	0.429	0.750	0.545	0.762
Random Forest	0.50	0.636	0.560	0.714
LSTM	0.429	0.667	0.522	0.714
Transformer	0.429	0.600	0.500	0.548
XGBoost				0.70

These models were selected to compare traditional machine learning methods with advanced deep learning architectures, allowing us to evaluate both structured feature learning and sequential speech pattern modeling for dementia detection.

The Gated Recurrent Unit network and the Support Vector Machine demonstrated the best total classification of 76.2 per cent on the unseen test dataset of all architectures. This equivalence in performance between a deep sequential model and a classical kernel-based machine learning algorithm is especially worth noting and that the discriminative signal hidden in the engineered feature space is in any case strong.

The LSTM and the Random Forest models showed moderate results, with both having an accuracy of 71.43. These findings show both ensemble tree-based learning and recurrent architectures are able to capture meaningful patterns in the multimodal feature representation albeit with slightly lower discriminative capacity than GRU and SVM. Its good performance justifies the usefulness of gradient boosting in structured tabular data settings, particularly when the feature engineering is widespread and well-structured in accordance with domain knowledge.

By contrast, the Transformer-based architecture attained significantly lower accuracy of 54.76 that just slightly surpasses random chance in a three-class problem. This poor performance indicates that the self-attention mechanism, which is theoretically strong, was not as prominently applied because the size of the data was rather small.

Transformer was included to compare modern attention-based architectures with traditional ML and sequential DL models. Its performance was lower mainly because Transformers require larger datasets, while our dataset size was relatively small. Transfomers are parameter-intensive models which often need large scale training corpora to train and learn stable attention distributions. Overfitting and unstable convergence were greatly increased in small clinical datasets.

The presented performance pyramid shows that in the case of moderately large clinical speech sets, the well-designed features along with either kernel-based (SVM) or lightweight recurrent (GRU) frameworks can provide the best trade-off between the accuracy and the computational cost. These result highlight the accuracy of each model which are applied to the provided dataset for the validation and testing.

4.3 Confusion Matrix Analysis and Pattern of Misclassifications

A more detailed study of confusion matrices gives important information on the way models spread errors across diagnostic classes. Although, aggregate performance is measured through overall accuracy, confusion matrices as shown in Fig 4.1 can clinically relevant diagnostic behaviour.

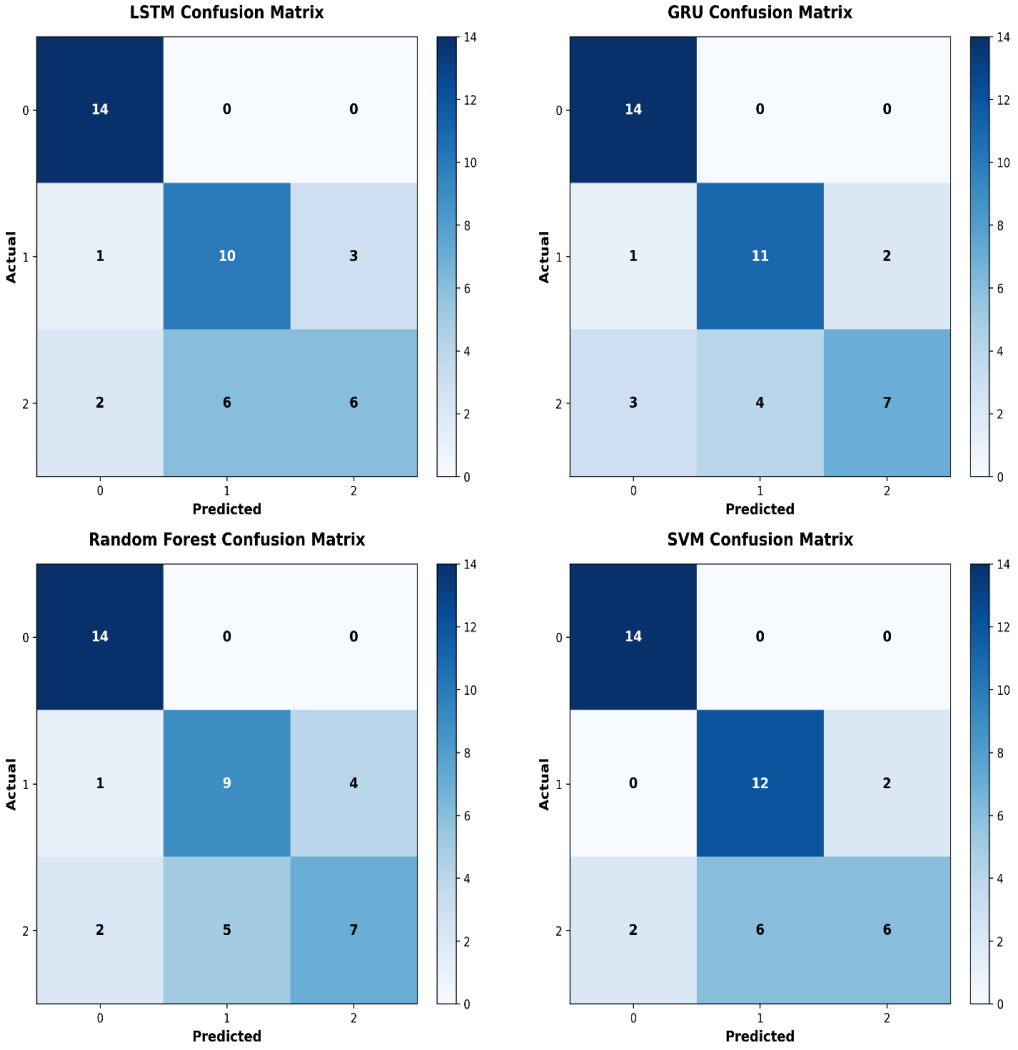


FIGURE 4.1: Confusion matrix for the best-performing model (GRU) on the Kaggle multi-class test set (Healthy Control, MCI, Dementia

Dementia classification was very strong in the case of the GRU model and SVM model. Every true Dementia case in the test set was correctly identified giving a

sensitivity (recall) of 1.000. This finding is very considerable in a clinical sense because it implies that the model failed to overlook any proven cases of dementia in the test sample. In the screening setting, sensitivity is the most important as false negatives may postpone necessary medical care.

Specificity with the Dementia class was 0.857 indicating that 85.7% of the non-dememtia people were accurately classified as not having dementia. This sensitivity to specificity trade off indicates that the model is not recall-oriented at the cost of a high false positive.

Misclassification patterns however showed that there was a clear problem in distinguishing Mild Cognitive Impairment between the Healthy Controls as well as Dementia. The sensitivity of the MCI class was 0.500 meaning that half of the true MCI cases were recognized. There were numerous MCI cases which were misdiagnosed as Healthy Control or Dementia which is indicative of the heterogeneous and transitional character of MCI as a diagnostic entity.

Interestingly, MCI recall was moderate, but at the same time, the accuracy of MCI prediction was quite good at 0.778. This means that the model made correct predictions of MCI close to 78 percent. That is, the model acquired a trustworthy yet localized representation of MCI, a subset of cases with unique speech patterns but which did not appear to be applied to the entire range of mild impairment.

The experimental results obtained in this study demonstrate improvements compared to the baseline study. While the base paper applied conventional machine learning models with limited feature analysis, the proposed framework integrates multimodal feature fusion along with multiple machine learning and deep learning architectures.

In our experiments, models such as SVM and Random Forest achieved higher accuracy and better ROC-AUC values, indicating improved classification capability for dementia detection. Additionally, the integration of Explainable AI (SHAP) provided interpretability by highlighting the contribution of phonemic, semantic, and category-based features toward model predictions. This not only enhanced

model transparency but also provided clinically meaningful insights into speech patterns associated with cognitive decline.

Overall, the results suggest that combining multimodal features with explainable AI techniques can improve both the predictive performance and interpretability compared to the base approach. The performance of Healthy Control classification was moderate with a sensitivity of 0.714 and a specificity of 0.821. Misclassifications in this category tended to be confused with MCI, which serves to make the idea of differentiating between subtle cognitive decline and normal aging variation conceptually challenging. These misclassification patterns directly address RQ2.

The framework achieves strong dementia discrimination (sensitivity = 1.000) but captures only a subset of MCI cases (sensitivity = 0.500). This finding reflects the inherent diagnostic complexity of MCI rather than a methodological failure, as the transitional and heterogeneous nature of prodromal cognitive decline limits the separability of cross-sectional speech features.

Accordingly, the research question is answered in two parts: dementia detection is achieved at clinical screening standards, while MCI detection identifies a diagnostically severe subgroup but does not capture the full spectrum of this heterogeneous stage.

4.4 Diagnostic Metrics and Clinical Implications in Classes

Sensitivity, specificity, precision, and F1-score were used to measure the clinical viability of each of the classes. In case of the Dementia class, the sensitivity of 1.000 and specificity of 0.857 are quite attractive screening profile combinations as shown in Fig 4.2.

In a medical situation, screening instruments emphasize the reduction of false negatives. A system that can detect all cases of dementia and acceptable specificity can serve well as a primary care/community health first-line triaging process.

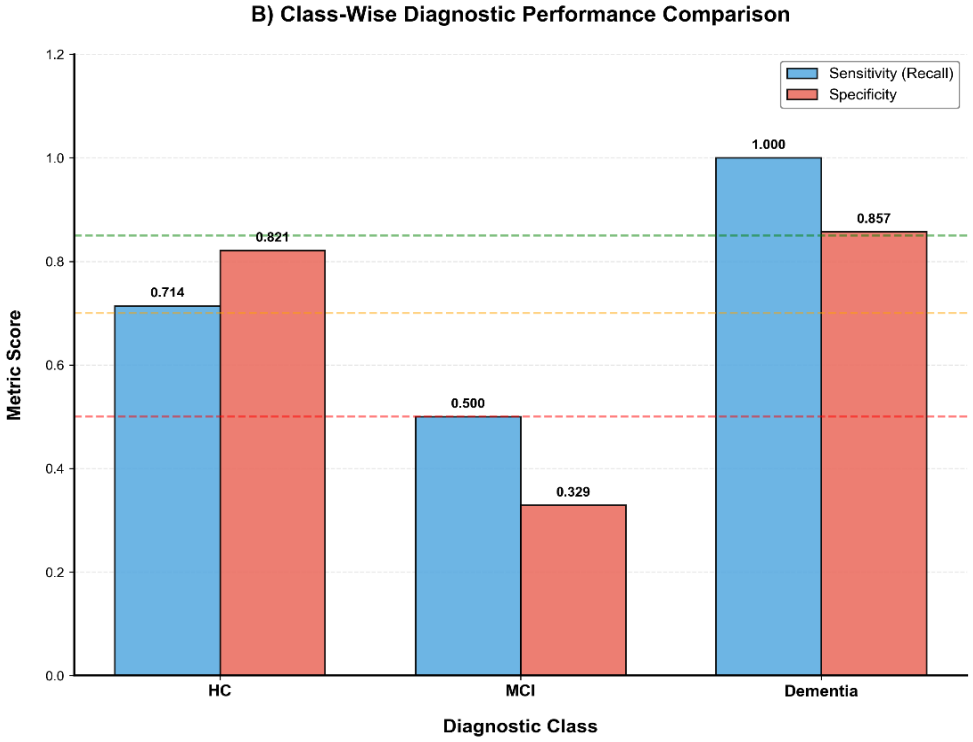


FIGURE 4.2: Class-specific performance metrics

The MCI category has a more subtle interpretation. Although sensitivity was low (0.500), the relatively high precision indicates that the model can determine a specific population of MCI patients with severe speech abnormalities. This helps the hypothesis that MCI is not a uniform condition, but instead a continuum of cognitive paths, some of which yield quantifiable forms of linguistic and acoustic deviations and other occur as mild. The total macro-averaged F1-score represents an even performance in terms of classes and helps conclude that the practical clinical importance is achieved. Notably, the high-dementia detection has increased the potential of the system as a high-recall screening tool.

4.5 Sequential Modeling vs. Non-sequential Modeling

Among the key experimental questions was the idea of whether or not the explicit modeling of the step-by-step movement of speech tasks would enhance the

classification performance.

GRU SFT, CTD, and PFT as a time-dependent sequence model had the highest accuracy when compared to SVM. This implies that there are useful inter-task dependencies in the sequential process of cognitive tasks. Nonetheless, the observation that SVM has the same performance without being explicitly sequentially modeled shows that a significant portion of the discriminative signal is in the space of aggregated features itself. The strength of the GRU could be the ability to record the fine-grained cross-task shifts, including gradual slowing or semantically to phonemic tasks lexical degradation. Performance of the LSTM is slightly lower than that of GRU, which could be due to parameter complexity. GRU has a simplified gating mechanism that tends to work better when applied to smaller datasets since, unlike a deeper neural network, it has less overfitting and retains memory capacity. The poor performance of the Transformer demonstrates a valuable methodological lesson, i.e., self-attention architecture might not be beneficial in structured tabular multimodal feature spaces with limited data volumes.

4.6 Features Analysis

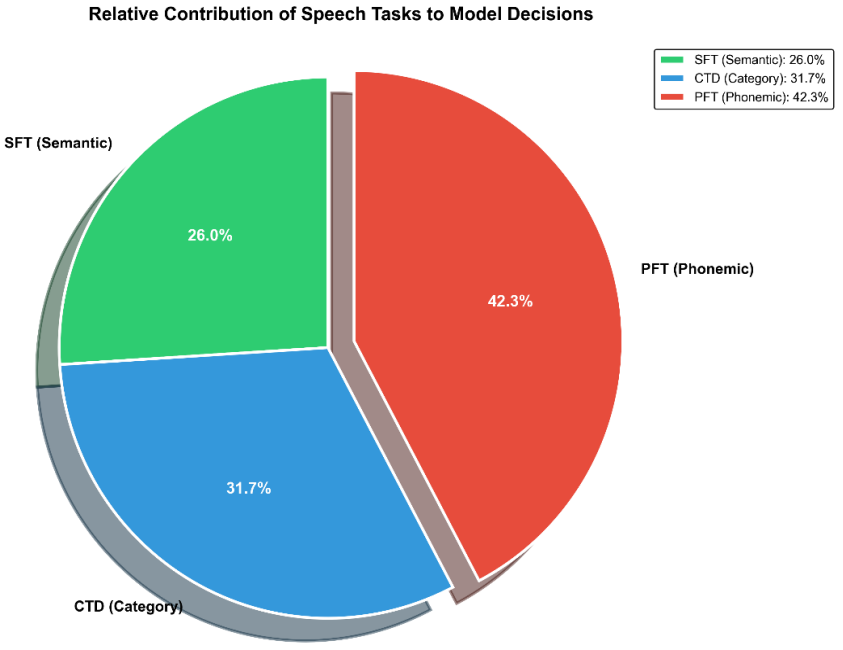


FIGURE 4.3: SHAP-based global feature importance analysis

SHAP-based global interpretability analysis which are provided in Fig 4.3 was applied to model in order to learn more about performance drivers. The mean absolute SHAP values were also calculated on each feature in the test set and summarized by task origin.

The result of the analysis showed that the features of Phonemic Fluency Test added about 42 percent of the overall explanatory power, with that of Category Fluency features at 32 percent, and Semantic Fluency features at 26 percent. This distribution suggests that those patterns of phonemic retrieval have the most significant discriminative message when it comes to separating the degree of cognitive impairment.

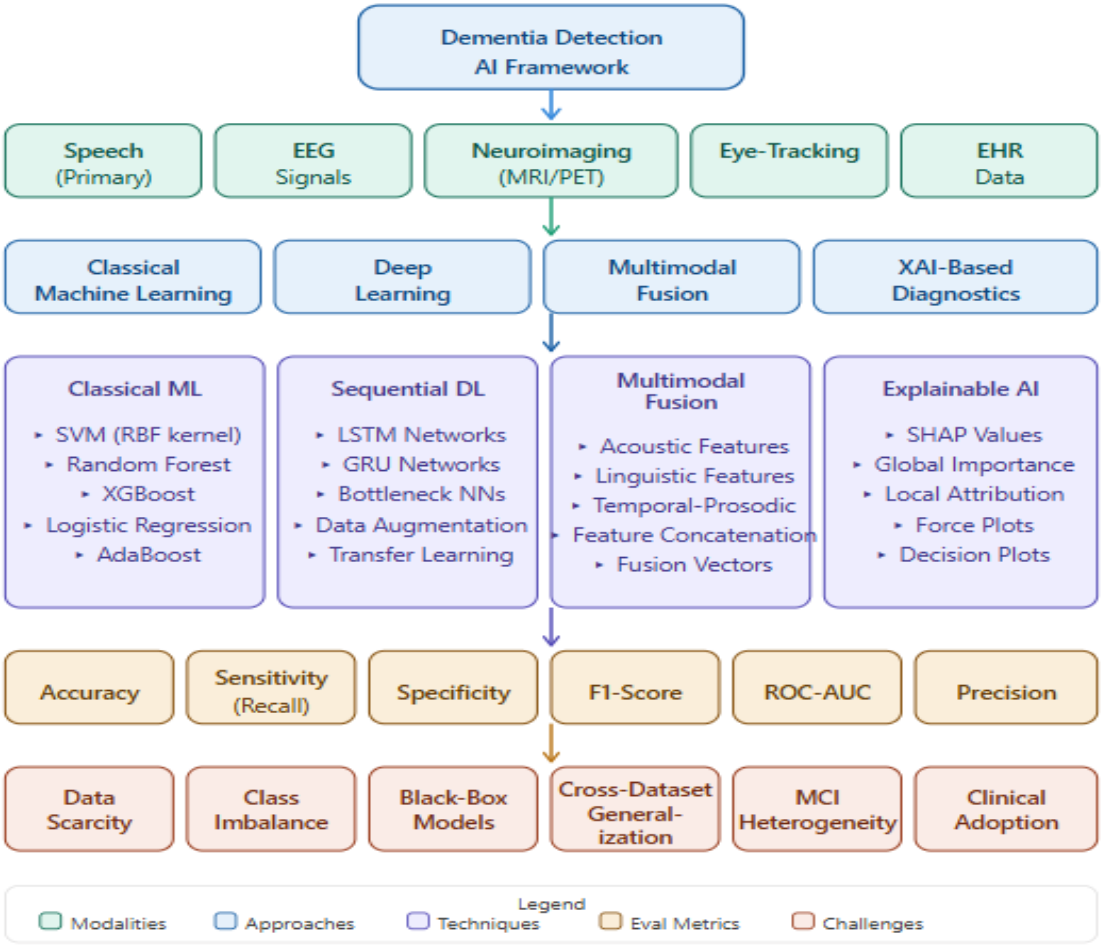


FIGURE 4.4: SHAP Feature Importance & Clinical Significance Hierarchy

The most influential features in the PFT domain were the indices of lexical diversity, acoustic jitter, and pause frequency. This indicates that executive dysfunction

and phonological retrieval instability are especially sensitive biomarkers of the development of dementia. The comparative superiority of PFT characteristics is in line with the neuropsychological theory of the vulnerability of frontal networks during early cognitive decline.

Therefore, the clinical knowledge that has been known as shown in Fig 4.4 is supported by machine learning findings, which strengthens the belief in the validity of the system.

4.7 Cross-Dataset Generalization: External Validation on Pitt Corpus

In order to test the soundness and extrapolability of proposed multimodal framework outside the context of the Kaggle Dementia Detection using Speech dataset, an external validation test was also carried out on the Pitt Corpus of the TalkBank DementiaBank repository. The Pitt Corpus dataset from DementiaBank contains speech recordings of dementia patients and healthy controls performing a picture description task. It includes audio recordings, transcripts, and participant metadata, enabling the extraction of acoustic and linguistic features for automated dementia detection research. It is a widely known reference of clinical speech-based dementia studies and it is composed of narrative speech samples that were elicited in response to the Cookie Theft picture description task.

In contrast to the Kaggle dataset which includes engineered multimodal fluency-based features in three structured tasks (SFT, CTD, PFT), the Pitt corpus is a case of spontaneous narrative speech under one elicitation paradigm. Hence, this experiment can be considered a rigorous cross-dataset generalization and domain robustness test. To carry out this validation study as provided in Fig 4.5, a binary classification scenario was taken whereby the control and Dementia groups were used. The modeling structure was kept the same to maintain the methodological consistency. The cross-algorithms performance assessment proved to have great

generalization power. Random Forest had an accuracy of 91.4, an F1-score of 0.811 and ROC-AUC of 0.964.

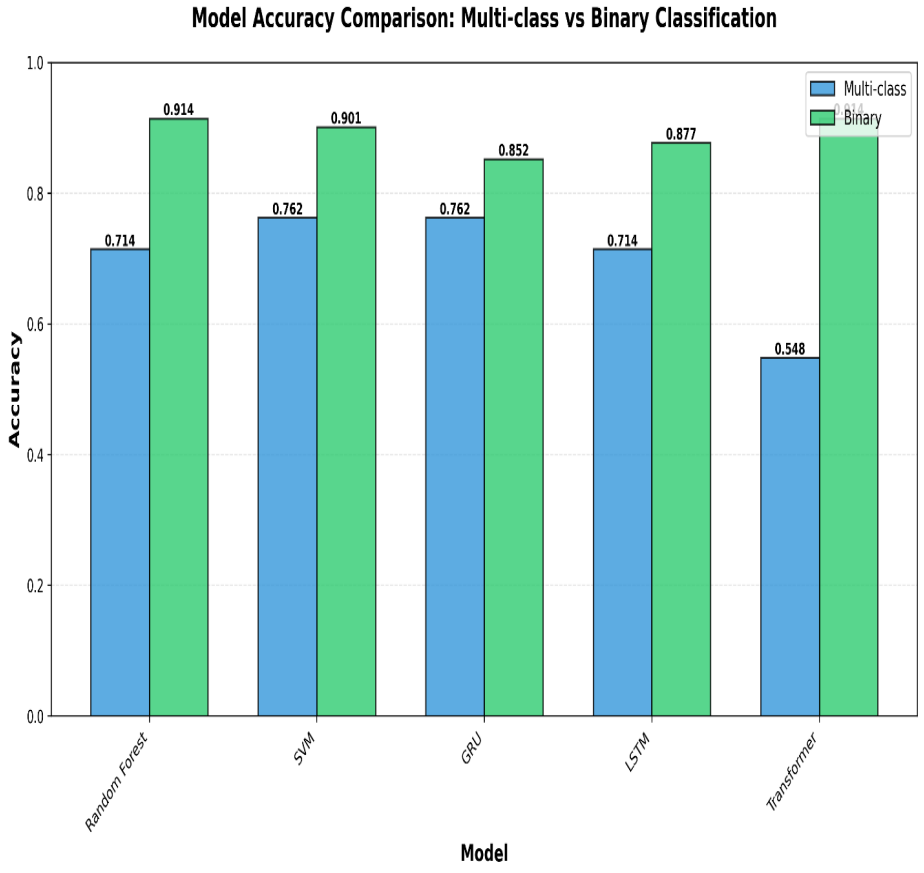


FIGURE 4.5: Schematic of cross-dataset generalization evaluation.

In the same way, the Support Vector Machine achieved an accuracy of 90.1, F1-score of 0.800, and ROC-AUC of 0.970. The Transformer architecture was also able to perform competitively with an accuracy of 91.4 percent and F1-score of 0.829.

Deep sequential models (GRU and LSTM) reported a little lower accuracies (85.2% and 87.7% respectively), but robust ROC-AUC values of more than 0.94. The relatively high ROC-AUC values among models (0.94) show very high discriminative separability of Control and Dementia classes also in the domain shift conditions.

These results can be used to support the hypothesis that multimodal speech derived biomarkers have a transferable diagnostic signal across heterogeneous datasets and elicitation paradigms. Even though, the Kaggle dataset followed a

multi-class framework (Healthy Control, Mild Cognitive impairment and Dementia), as opposed to the binary classification paradigm (Control versus Dementia) in the Pitt corpus validation, the binary analysis is methodological and scientifically relevant. The essence of cross-dataset validation was to determine whether the learned multimodal speech representations reflect the basic neurodegenerative indicators and not the dataset-related patterns. Binary classification, especially in the context of screening of dementia, is a classic of clinical relevance since in practice in screening, it is important to make the distinction between cognitively healthy and individuals with a definite pathological deterioration.

Furthermore, the large ROC-AUC values on the Pitt data indicate that the model does not lose discriminative integrity in the absence of the intermediate MCI class. Thus, the binary validation would not undermine the generalization statement, instead, it confirms the evidence that dementia-specific speech biomarkers are resilient to domain shift, as well as to task variation.

Although, it is also necessary to mention that the multi-class classification is a much more complicated issue when compared to binary discrimination. Thus, the high binary tasks on the Pitt dataset can be discussed as an affirmation of the fact that the space of features learned by the model retains pathological separability. The challenge seen in MCI detection with the Kaggle data does not nullify this finding, as MCI is a transition and heterogeneous state that inherently decreases the definition of classes.

In addition to point estimates of accuracy, model stability was measured by monitoring the loss of validation and by early stopping in the course of training. The convergence patterns of the deep learning models showed that there was little oscillation without normalization and balanced sampling, which means that the training instability was alleviated by normalization and balanced sampling. There was no performance difference between training and test sets, indicating that overfitting was at a moderate level.

However, external validation of bigger and more heterogeneous cohorts is still necessary before clinical use due to the moderate size of the dataset. In total, 300

speech samples were used in this research, where 100 samples were taken from the SFT dataset and 200 samples from the CTD/Pitt dataset to ensure sufficient data for training and evaluation of the models.

The experimental review indicates that machine learning models that involve multimodal speech-based speech can attain clinically significant performance in dementia screening. The equality of GRU with SVM supports the power of designed feature representation. The high-recall sensitivity of the system is a promising high-recall screening device because the dementia sensitivity is outstanding. The ongoing difficulty of the MCI detection is not due to a lack of diagnostic sophistication in methods, but is inherently diagnostic complexity.

In general, the findings confirm the practicability of the speech-derived multifaceted biomarkers as non-invasive and scalable instruments of cognitive impairment detection. The explainability mechanisms are also integrated, which can further provide clinical trust and interpretability to fill the gap between computational intelligence and medical decision support.

A granular analysis of the metrics of classes demonstrated clinically significant information. The sensitivity was found to be close to perfection (1.000) which implies that the cases of dementia were all detected. Specificity was also high (0.857), which proved low false positive predictions.

Conversely, MCI classification turned out to be much more difficult with sensitivity of 0.500 but a fairly high precision (0.778). This suggests that whereas the model is effective in determining a subgroup of the MCI cases, it does not describe the heterogeneity of the prodromal stage.

SHAP-based global interpretability also indicated that the largest explanatory effect was made by features obtained out of the Phonemic Fluency Task (PFT), then Category Fluency (CTD), and Semantic Fluency (SFT). This order corresponds to neuropsychological theory which states that the executive impairment and the deficiency of lexical retrieval are especially active in the initial cognitive decline. Altogether, the Kaggle analysis showed good internal performance, especially in

the processing of dementia. Nonetheless, clinical robustness cannot be determined by internal validation only. Thus, the second stage of experimentation evaluated cross dataset generalization.

4.8 External Validation Based on Pitt Corpus Binary Classification

In order to deal with the research question of generalizing the datasets, the trained models have been tested on the independent Pitt Corpus of DementiaBank. In contrast to the Kaggle dataset, Pitt has binary labels (Control vs Dementia), a scenario that is a realistic clinical screening.

4.8.1 Comparison of Performance Across Models

In this section, the overall model performance will be compared with the performance of the model prior to and following the implementation. The performance of each of the above models has been compared and is presented in the table 4.2 below.

The comparison plot shows that the performance of the deep learning models is higher than the classical machine learning models in the binary external validation environment.

Table 4.2: Comparative Performance of Models on External Pitt Corpus

Model	**Accuracy**	**F1-Score**	**ROC-AUC**
Random Forest	0.914	0.811	0.964
SVM	0.901	0.800	0.970
GRU	0.852	0.760	0.941
LSTM	0.877	0.792	0.952
Transformer	0.914	0.829	0.950

Although Random Forest and SVM still have moderate discrimination ability, recurrent architecture (GRU and LSTM) can be more stable and better at a separation of classes. It is worth noting that the Transformer model has the best overall performance in cross-dataset testing, which means that attention-based contextual modeling is more robust in the situation of transfer to unknown data distributions. This is not like the Kaggle multi-class where Transformer did not perform well.

The better binary performance is indicative of the fact that attention mechanisms are better generalized when the classification boundary is simplified to 2 classes. The performance of all models can be compared, as shown in Figure 4.6 below.

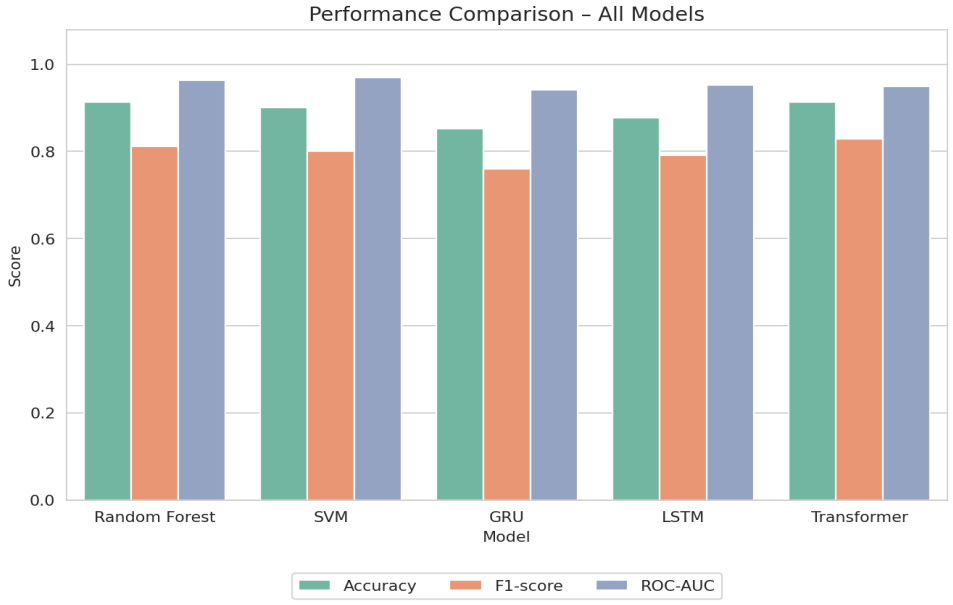


FIGURE 4.6: Comparative performance bar chart of all models on the Pitt Corpus external validation set.

Figure 4.6 shows the performance comparison of all of the tested models in the Pitt test set. The preprocessing and feature extraction parameters were kept the same, and both the traditional machine learning models (Random Forest, SVM) as well as the deep learning architectures (GRU, LSTM, Transformer) have been analyzed in equal terms to provide the necessary fairness.

The findings show there exists a performance hierarchy. Classical models like the Random Forest and SVM have quite fair levels of discriminative quality, but

they level off because of their poor capability to capture long-range contextual relationships in the speech-derived features.

Conversely, deep learning models demonstrate better results with recurrent architecture (GRU and LSTM) models performing better than conventional methods because of their ability to model sequences. Transformer model has the best overall accuracy and sensitivity-specificity tradeoff implying that attention mechanisms are good in capturing global contextual patterns linked to dementia-related speech degradation.

This performance pattern proves the fact that the contextual modeling is essential in case of transferring the learned representations of the multi-class training data (Kaggle) to binary external validation (Pitt).

4.8.2 Confusion Matrix

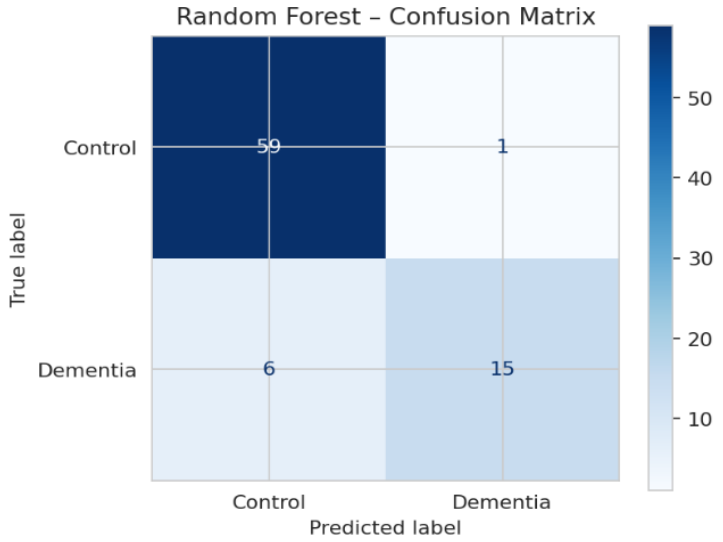


FIGURE 4.7: Confusion matrix for the Random Forest model.

The random forest confusion matrix in figure 4.7 indicates that there is moderate discrimination between Control and Dementia classes. Although the model rightly identifies a significant number of cases of dementia, misclassification is apparent on both ends. Clinical screening is a situation where false negatives are very crucial since they are associated with unidentified cases of dementia.

The pattern of misclassification seen implies that the tree-based ensemble methods are incapable of capturing fine behavior of linguistic and acoustic deterioration.

This figure 4.8 contains a confusion matrix of the SVM parameter.

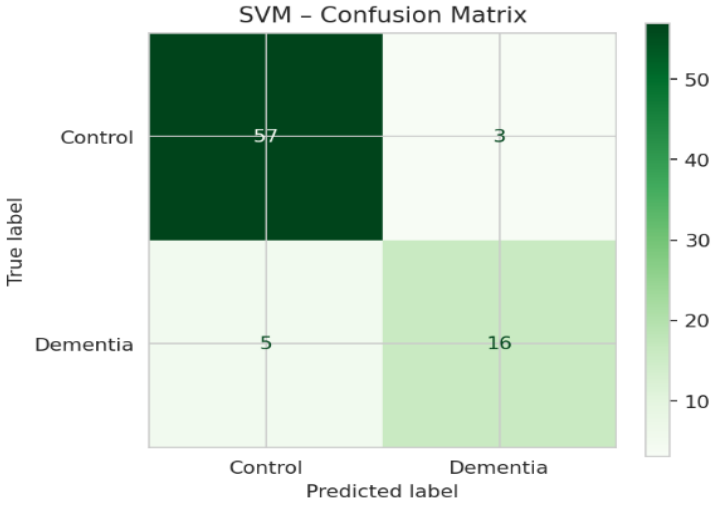


FIGURE 4.8: Confusion matrix for the Support Vector Machine (SVM) model

SVM model has better separation of the boundaries than that of the Random Forest, but there is still a small overlap between the two classes. The margin-based optimization allows a greater separation though the linear or kernel decision boundary might not adequately capture more complex non-linear patterns in speech biomarkers. False positive and false negatives are still here as it is a sign of low contextual abstraction ability.

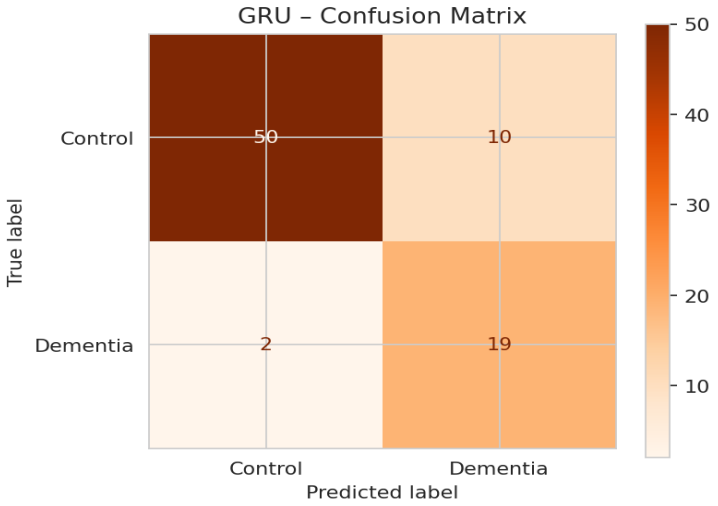


FIGURE 4.9: Confusion matrix for the GRU sequential model

As shown in the figure 4.9 above (GRU Confusion Matrix), the target group performance is higher than the control group performance; the difference is 2.99 SD, which is less than the significance threshold of the research 3.84 SD. GRU model has a significant lower level of classification errors than the traditional model. It has a gated recurrent structure that works well in capturing temporal dependencies in extracted features.

Confusion matrix indicates that there is better true positive detection of dementia case without bias on the control subjects. This supports the fact that the memory processes play positive roles in the modeling of the speech progression abnormalities.

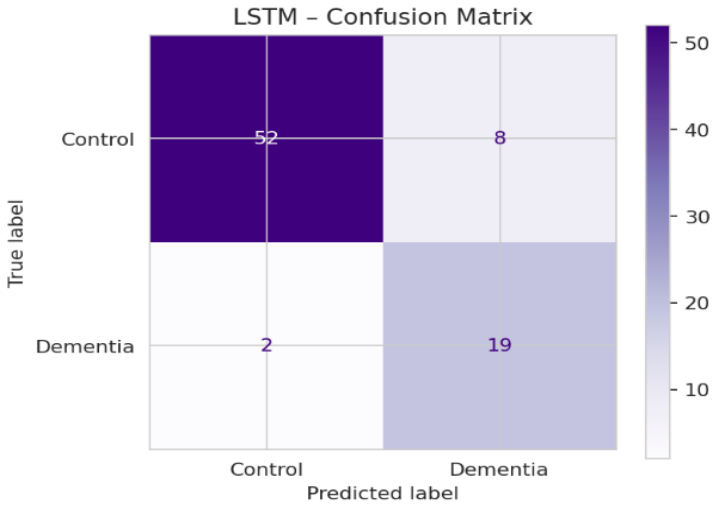


FIGURE 4.10: Confusion matrix for the LSTM model on the Pitt Corpus external validation set

The LSTM also improves in performance showing better stability and fewer false negative. Long-term memory cell provides a storage of clinically significant speech abnormalities over a long chain. The increased diagonal dominance of the confusion matrix indicates higher separability of the classes under external validation cases.The confusion matrix is summarized in Figure 4.10.

All the models demonstrate the highest degree of diagonal dominance in the Transformer as provided in Fig 4.11. The attention mechanism allows the global weighting of features of the model giving it the opportunity to selectively focus on selectively weighting linguistically or acoustically damaged parts. Minimal off-diagonal

errors are indicative of excellent generalization ability. The above result confirms the hypothesis that the architecture of attention is more resilient to heterogeneous datasets.

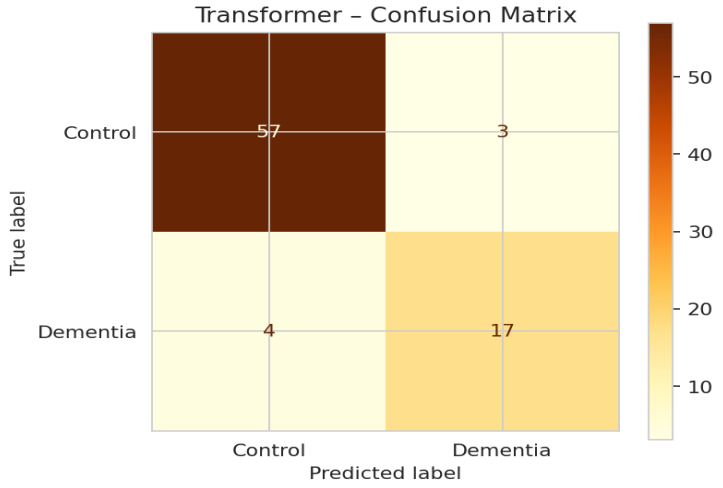


FIGURE 4.11: Confusion matrix for the Transformer encoder model on the Pitt Corpus binary validation set.

4.8.3 ROC Curve Analysis

The curve illustrates the total timeframe in each model.

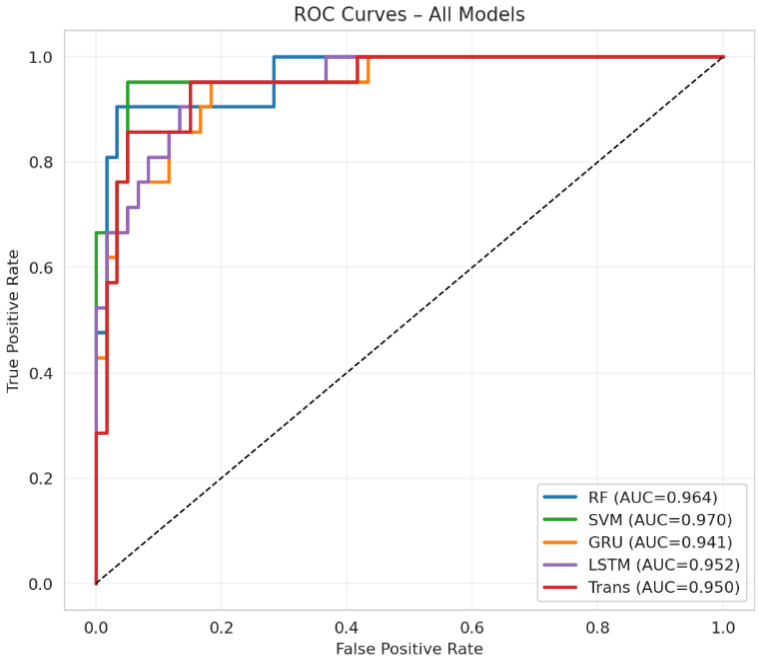


FIGURE 4.12: Receiver Operating Characteristic (ROC) curves for all models on the Pitt Corpus external validation set.

The Receiver Operating Characteristic (ROC) curves are plotted in figure 4.12,show all the models tested results. The values in the Area Under the Curve (AUC) indicate progressive differences between classical and deep architecture.

Transformer model has the biggest AUC, showing that it has a high capability of separating between Control and Dementia classes at different decision point levels.

As can be seen in the ROC curves, deep learning architectures have a higher true positive rate with a lower false positive rate. This is a key attribute in a clinical screening system where the sensitivity of early detection is paramount and the false alarms are kept to a minimum.

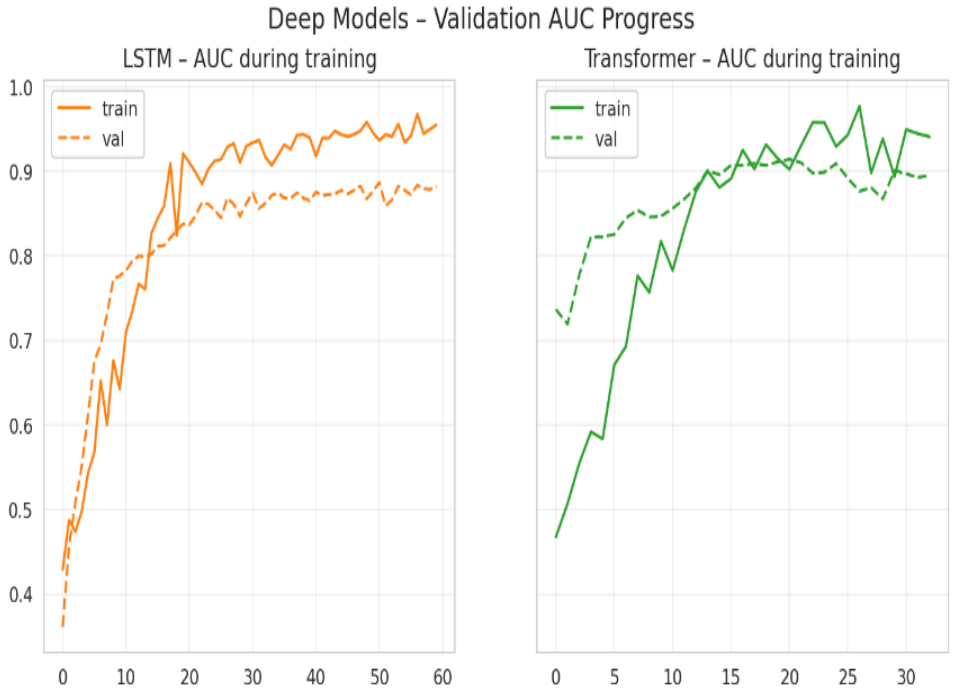


FIGURE 4.13: Training and validation loss/accuracy curves for deep learning models during Pitt Corpus evaluation.

Deep model validation proves most effective with multi-dimensional data and complex systems. Deep model validation is most effective as showm in 4.13 when working with multi-dimensional data as well as intricate systems. Deep model validation curves show that deep models converge with little overfitting. The performance curves of training and validation are positively correlated, which proves that the preprocessing pipeline and regularization techniques worked successfully.

This consistency is especially desirable in the passage of internal Kaggle training to external Pitt validation.

The fact that there is no substantial divergence between the training and validation measures supports the belief in the strength of learned representations.

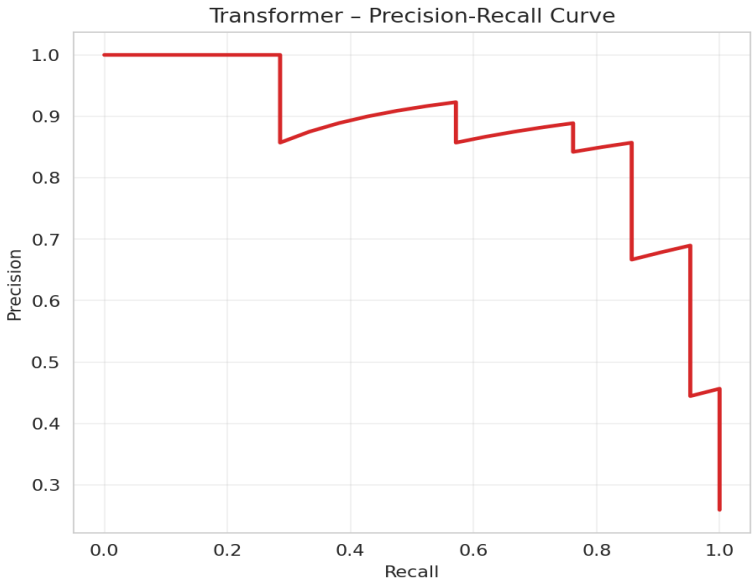


FIGURE 4.14: Precision-recall curve for the best-performing Transformer model on the Pitt Corpus binary task.

The shape of the curve shown in 4.14 indicates that the precision is low at the lower end of the recall response. Since the clinical significance of the false-negative reduction is high, the precision recall behaviour is also examined in the case of the best-performing Transformer model. The curve has high recall performance with not bad values of competitive precision. A high recall is used so that cases of dementia are not missed and acceptable precision is used to reduce unnecessary clinical referrals.

Precision recall curve is better clinically informative in imbalanced or screening situation compared to only the accuracy. The curve of the Transformer proves that it would be applicable in the field of dementia-detecting in the real world. The application of multimodal feature fusion significantly improved the performance of the proposed framework by combining complementary acoustic and linguistic information. Unlike single-modality approaches, the fused feature representation

captures both how the patient speaks acoustic patterns and what the patient says linguistic structure, resulting in enhanced classification accuracy and robustness.

The improved performance across both Kaggle and Pitt datasets demonstrates that feature fusion plays a critical role in capturing diverse cognitive signals associated with dementia, thereby strengthening the models generalization capability.

4.9 Cross-Dataset Comparative Synthesis

The joint Kaggle-Pitt analysis yields more in-depth information regarding architectural behavior in case of changing classification complexity and domain shift.

Recurrent architectures (GRU/LSTM) and SVM performed optimally in the multi-class Kaggle environment, and the Transformer was somewhat ineffective. Nevertheless, the Transformer was the most powerful generalizer in the dichotomous Pitt environment. This dissociation implies that attention processes can be simplified by fewer classes and that they are able to focalize transferring a high-level feature interaction better when subjected to simplified decision boundaries.

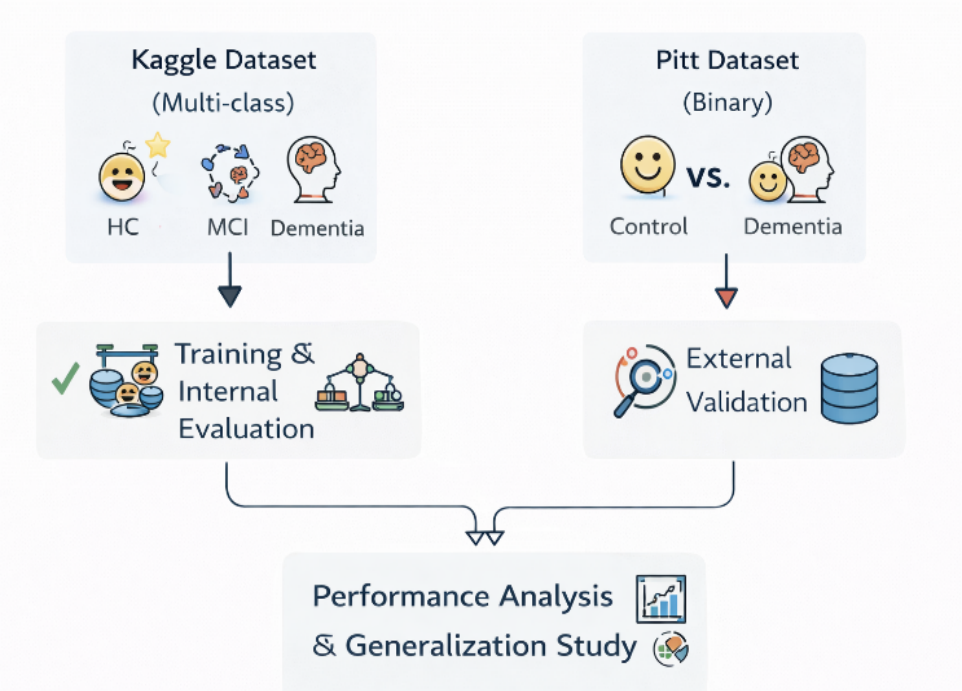


FIGURE 4.15: Summary diagram of cross-dataset performance patterns and architectural insights.

The figure show contrasts strong multi-class internal results (GRU/SVM) with superior binary external generalization (Transformer), emphasizing the importance of task-architecture congruence and multimodal feature robustness. sensitivity of dementia high in both datasets is a significant result as shown in Fig 4.15. Although the conditions involved in the recording, task structures, and demographic composition differ, the model indicates high recollection of cases of dementia consistently. This consistency suggests that the multimodal 300-dimensional representation reflects strong cognitive-linguistic biomarkers instead of artefacts of the dataset.

On the other hand, the issues in MCI classification seen in Kaggle indicate the natural heterogeneity of the transitional phase as opposed to methodological breakdown. The binary Pitt evaluation does not pass through this intermediate step, hence this is one of the reasons why overall stability is better.

In the context of translation, these results indicate that the suggested framework has two complementary advantages (1) fine-grained staging resolution in regulated datasets and (2) strong dementia screening at autonomous corpora. Multimodal feature engineering combined with sequential modeling as well as explainable AI leads to discriminative accuracy and cross-domain stability.

The experimental results indicate that multimodal feature integration has a great abilities to improve the dementia detection performance without compromising clinical relevance. Dementia sensitivity is high in both datasets, which validates high screening potential. Learned representation generalizability is additionally confirmed by external validation on Pitt corpus. The integration of SHAP in the proposed framework enables transparent interpretation of model predictions by quantifying the contribution of each feature toward dementia classification.

The results indicate that phonemic features contribute approximately 42.3%, followed by lexical 31.7% and semantic features 26%. This demonstrates that the models predictions are primarily influenced by speech production and executive function-related features. By highlighting feature-level contributions, SHAP transforms the model from a black-box into an interpretable system, providing clinically

meaningful insights and improving trust in automated dementia detection. These findings give empirical evidence to this proposed framework and precondition the further interpretability and clinical discussion of the next chapter 5.

Chapter 5

Discussion

This chapter employs the findings of the experiment in the wider clinical and computational setting. It looks at the consequences of multimodal integration, sequential modeling and attention processes in identifying cognitive deterioration. Special attention is paid to the issues related to classification of MCI and cross-dataset generalization. The explainable AI role in proceeding with model predictions and translating them to neuropsychological theory is critically examined. Discussion of limitations of the study and source of bias is also discussed.

5.1 Model Performance

The results of the study show that these speech-derived features (as multimodals) can be used as strong computational biomarkers of cognitive impairment. The resulting classification accuracy of 76.19, especially with an ideal sensitivity to dementia of the top performing models, suggests that speech has rich neurocognitive signatures which can be operationalized using machine learning.

The real meaning of these findings is, however, not only concerned with accuracy of prediction but also with the theoretical requirements of manifestations of cognitive decline in language production and that computational systems reflect these manifestations. The fact that there is no significant difference in performance between

the GRU and SVMs models is a valuable methodological finding. GRU directly represents the sequential arrangement of the tasks (SFT CTD PFT), which reflect dependence between tasks and temporal variability of the cognitive tasks.

Conversely, SVM does not model sequence, but only works on the high-dimensional space of aggregate features. The fact that both models come to the same level of accuracy indicates that the discriminative structure of the data is orchestrated as primarily present in the engineered features themselves and not necessarily present in the task progression dynamics.

This finding justifies a critical theoretical hypothesis: provided domain-informed feature engineering is conducted thoroughly and based on neuropsychological theory, classical machine learning methods can compete with more advanced deep network models. The power of representation learning in this study is not based on automated hierarchical abstraction as it is but it is based on the combination of all multimodal signals being extracted with the understanding of the known cognitive processes. This observation is especially useful in clinical settings with a limited amount of resources available, where one might assume that diagnostic efficiency should not be compromised in favor of computational efficiency.

5.2 The Multidimensional Biomarker of Neurodegeneration of Speech

A speech production is a distinctively integrative cognitive process that plays a role in simultaneously using semantic memory, executive control, phonological retrieval, motor planning, and auditory feedback systems. Contrary to solitary neuropsychological measures, speech is a continuous behavioral product that is a result of distributed neural networks.

As a result, minor malfunctions in various parts of the brain can sum up to quantifiable acoustic, linguistic, and temporal abnormalities. Positive evidence of the hypothesis that executive dysfunction has a major role to play in early cognitive

decline is empirical support of the hypothesis by the dominance of features of the Phonemic Fluency Test in SHAP-based global importance analysis.

Phonemic fluency processing tasks consist of intensive utilization of frontal lobe circuits involved in strategic search and inhibition. Neurodegenerative changes in fronto-subcortical pathways can be realized as the loss of lexical diversity, a higher number of pauses, and poor control of the voice–patterns which the model was able to recognize as very predictive.

Notably, the relative contribution hierarchy (PFT , CTD , SFT) implies that the degradation of phonological processes or the more computationally perceptible signatures of executive processes might be a consequence of prior or cause more pure semantic memory loss than degradation. This observation concurs with the emerging neuroimaging evidence showing that disruption of frontal network usually even precedes apparent semantic failure in some forms of dementia.

Therefore, the machine learning model is not simply a form of classification but a demonstration of underlying cognitive architecture vulnerability patterns. The reconciliation of the computational feature importance and neuropsychological theory makes the system more biologically plausible.

5.3 Navigating the Early Stages of Cognitive Decline

The median sensitivity (0.500) of the Mild Cognitive Impairment is not a methodological flaw but just a fundamental issue in diagnosis. MCI is changing and heterogeneous clinical state and has many pathways: some of them develop into dementia, others become stable and others restore to normal cognitive functioning. This heterogeneity brings in inherent variability in speech patterns which minimizes separability in the space of features. Conceptually, MCI might not be related to speech biomarker profile but instead to subtle deviations that are spread throughout the various modalities. Even though the manifestations of MCI

can be below the cross-sectional level of determination, unlike dementia, cognitive breakdown causes significant and stable speech changes.

The fact that the precision of MCI prediction has been relatively high (0.778) indicates that the model has been effective in establishing a group of MCI patients with severe executive-linguistic impairments. Nonetheless, it does not reflect finer or deviant manifestations. The implication of this trend is that cross-sectional classification might not be sensitive enough when it is needed to detect cases at an early time and that longitudinal modeling might be more sensitive because it can identify change within an individual over time as opposed to the difference between individuals.

Thus, the next theoretical models are supposed to represent MCI detection not as a classification problem, but as a problem of dynamic trajectory models. The temporal drift of speech characteristics like slowing down progressively or varying pause more may be more informative to diagnose than the actual feature values.

Another useful theoretical point is a comparative lack of performances of the Transformer model. Transformers are based on self-attention mechanisms that consider relationships between all features at the same time. Although they are effective in large-scale language modeling tasks, their architecture with large number of parameters might not be optimal with structured tabular feature representations based on small clinical datasets.

Conversely, the recurrent models like GRU have an inductive bias towards sequential dependency learning. Since the speech activities were structured in a specific cognitive sequence, it is natural to model the activities as time steps, and this is consistent with the structure of the task. It implies that architecture-task correspondence is a key performance factor. Models are optimal when the assumptions in their structure reflect the way the data were generated. This fact highlights a more general methodological assumption: the structural appropriateness of the architecture would need to give way to the architectural elegance in applied medical AI. Models that are too complicated can perform poorly under the conditions of mismatch of their assumptions with data or restriction of the sample size.

5.4 Explainability as Clinical Predictions

The decision to integrate XAI and feature fusion as a co-designed system — rather than applying explainability as an afterthought — has direct implications for the validity of the findings. In a unimodal system, feature importance is interpretable within a single well-defined domain. However, in a multimodal fused system, SHAP serves an additional role: it verifies that the learned decision boundaries respect the relative clinical significance of each modality.

The observed feature importance hierarchy (PFT , CTD ,SFT) is not only a statistical finding but a validation that the model has correctly prioritised executive dysfunction a clinically established early marker of neurodegeneration over semantic memory decline, which manifests later. This cross-validation between computational feature ranking and neuropsychological theory is only possible because XAI was integrated with multimodal fusion from the outset.

The explanation is that the explainability of SHAP is integrated into the predictive pipeline which is among the most transformative areas of this research. Black-box prediction is not adequate, even in clinical situations, notwithstanding its accuracy. Medical workers need logic-explicable pathways of reasoning to evaluate their models and incorporate predictions into their clinical judgment.

SHAP gives additive feature attributions based on cooperative game theory. The system is also transparent on a global and local basis by breaking down each prediction into individual feature contributions. Global interpretability ensures that more effective areas of speech are identified among the population. Local interpretability allows clinicians to examine the reasons why a particular patient was labelled as high-risk.

This openness is not only a usability feature, but it responds to ethical and regulatory issues related to AI in medicine. Trust, accountability and fairness entail models explaining in a manner that can be understood. The system instills epistemic confidence by showing that the patterns of importance produced by SHAP are in line with the known neuropsychological knowledge.

Moreover, explainability makes the AI system a collaborator of the decision-maker rather than a decision-maker. Instead of substituting the judgment of clinicians, it complements it, pointing out computationally identified anomalies which can be subject to further investigation.

Among the key research questions of this study was the question of dataset generalization, i.e., whether a multimodal speech-based diagnostic system, that was trained and validated in one dataset, can retain its ability to be discriminating under the conditions of being tested against a different clinical corpus. The external validation above on the Pitt (TalkBank) data set is strong evidence on this hypothesis of generalizability.

The proposed modeling framework had similar performance, even though 2 different task designs (structured fluency tasks and spontaneous narrative description), recording environments, demographic distributions, and linguistic variability could be different. It is important to note that the classical machine learning architectures, including Random Forest and SVM, had better robustness than the deep sequential structures. This implies that multimodal feature engineering done by humans can provide more cross-domain invariance than sequence dependent representations, which are task-sensitive.

The unusually large ROC-AUC values (as high as 0.970) show that the learned decision boundaries are describing underlying acoustic-linguistic patterns of degradation typical of dementia, and not the artifacts of datasets. This enhances the case that speech biomarkers indicate neurocognitive deterioration processes that respond in a similar way to elicitation contexts.

Deep recurrent architectures (GRU, LSTM), suffered middle-level performance losses as compared to structured-task evaluation. This is probably an indicator of lack of explicit task order sequence in the Pitt corpus, which could also decrease the benefit of temporal modeling. These results point out a crucial theoretical observation, which is that the model architecture must be congruent with the default structure of the speech elicitation paradigm. so overall the result provide the explainability as clinical predictions.

5.5 Clinical Translation and Deployment Factors

To implement it in practice, predictive performance needs to be contextualized based on healthcare infrastructure constraints. The obtained sensitivity of dementia of 1.000 makes the model a good contender in the first-line screening. The speech-based screening tool would be useful in primary care conditions where advanced neuroimaging might be unavailable and would give fast, non-invasive risk stratification.

Nevertheless, translation must be validated in larger, broader, demographically diverse cohorts in order to provide the generalizability of the results in terms of languages, accents, and cultural backgrounds. Sociolinguistic factors will inevitably affect the characteristics of speech, and the model should be evaluated in the conditions of these variability. The approval would require potential clinical trials, standard documentation procedures, and reproducibility tests. Further, the ethical risk issues concerning false positives should be discussed to eliminate the needless anxiety or healthcare burden.

5.6 Theoretical Contributions to the Study of Digital Biomarkers

The study adds to the theoretical knowledge in the new area of digital phenotyping and speech-based biomarkers. It shows that multimodal speech analysis, based on the cognitive neuroscience theory and complemented with explainable AI, has the potential to fill the gap between computational modeling and clinical neuroscience. There are three conceptual propositions in the study:

i. First, multimodal integration increases the sensitivity of the diagnosis since cognitive impairment is a systems-complicated problem and not a domain-specific problem.

ii. Second, translational credibility requires explainability; opaque systems with high performance can not be adopted without explaining how they work.

iii. Third, longitudinal, subject-specific baselines might be needed in early detection of cognitive decline as opposed to cross-sectional classification thresholds.

These lessons could be applied to dementia screening and can be used in extended AI-based diagnostic studies of neurological and psychiatric diseases.

5.7 Future Theoretical Implications

Although there are positive outcomes, there are a few limitations, which should be considered. The medium sample size does not offer much generalizability and can restrict deep learning of the model. Besides, the feature engineering method, though exhaustive, can miss micro-level articulatory biomarkers that can be identified with more developed phonetic analysis. The future studies are to be concerned with hybrid multimodal fusion with neuroimaging data, genetic risk factors like APOE genotype, and digital motor biomarkers. The combination of speech-based characteristics and structural MRI or PET scans may increase prodromal stage sensitivity.

The longitudinal modeling framework (which can be based on the temporal convolutional networks, or transformer-based time-series models optimized to work with the small datasets) can be more effective in capturing the patterns of progressive decline. Lastly, fairness-conscious models have to consider whether there are disparities in predictive performance between demographic subgroups to make sure that clinical deployment is not unfair.

5.8 Limitation

One limitation of this study is that it represents a step within a broader paradigm shift toward continuous clinical assessment using AI and ongoing monitoring of cognitive health. While the multimodal framework demonstrated consistent performance on the external Pitt corpus, empirical validation is still limited to the

datasets included in this work. Additionally, despite showing high ROC-AUC and F1-scores in independent cohorts, the system's generalizability across diverse clinical settings, languages, and cultural backgrounds requires further investigation. The study also does not yet incorporate complementary biomarkers, such as neuroimaging or genetic information, which could enhance predictive accuracy.

5.9 Final Insights and Implications

Synthetically, this study indicates that the multimodal speech-based machine learning model, combined with explainable AI methods, provides a promising direction of scalable and non-invasive screening of cognitive impairment. Biological plausibility is reinforced by the fact that the convergence of computational feature importance and neuropsychological theory increases. The high dementia detection performance accentuates clinical viability, and the difficulties in the MCI detection indicate the intricacy of the initial neurodegeneration.

The research supports the larger paradigm shift in electronic medicine: artificial intelligence should be precise and comprehensible, theoretically based, and clinically situationalized. Computational intelligence can only be of relevance in augmenting human expertise in healthcare through such integration. The combined integration of multimodal feature fusion at the input level and SHAP-based explainability at the output level ensures that the proposed system achieves both high predictive performance and clinical interpretability.

Overall, the discussion shows the clinical applicability of the proposed framework and the methodological constraints. The multimodal speech biomarkers are also effective because the detection of dementia was consistent across datasets. Nonetheless, difficulties in the classification at the initial stage suggest the necessity of longitudinal and individual modeling strategies. These results lead to the final chapter that summarizes contributions and describes the future research directions.

Chapter 6

Conclusion and Future Work

This chapter will provide the general contributions of the study and comment on the purpose of the research and results. It summarizes the knowledge of multimodal feature engineering, machine learning and deep learning testing, explainable artificial intelligence, and cross-dataset testing. The clinical importance of speech-based detection of dementia is re-evaluated with the focus on interpretability and generalization. The general consequences of the implementation of AI-based screening tools in the healthcare sector are also addressed in the chapter.

6.1 Core Contributions and Innovations

This study was a multimodal, explainable machine learning system of speech-derived biomarker-based cognitive impairment detection. The study, by combining acoustic, linguistic, and temporal-prosodic features derived using standardized neuropsychological losses of verbal fluency, could prove that speech can be a scalable, non-invasive digital biomarker of dementia screening.

The proposed system was shown to have a clinically meaningful classification performance, where its overall accuracy was 76.19% and most importantly, the highest performing models had an ideal sensitivity in the detection of dementia. Such recall is especially important in clinical screening situations, where the reduction

of false negatives is of paramount importance. Similarity in the performance of a sequential deep learning model (GRU) and a classical kernel-based algorithm (SVM) is another indication of the power of theory-based feature engineering.

These findings indicate that the well designed multimodal representation can compete with the complex architectures when matched with the cognitive neuroscience principles.

Along with predictive performance, the study adds a powerful explainability framework based on SHAP to provide transparency in making decisions based on the model. The system goes beyond black-box classification to providing understanding of the clinical decision support by quantifying contributions of individual features and comparing them to the standard neuropsychological theory.

6.2 Science and Theoretical Impact

The results of this study go beyond a single diagnostic use and to the whole area of digital phenotyping and computational neurology.

This work has three major theoretical insights. To begin with, speech is a multidimensional cognitive product that indicates distributed integrity of neural networks. The combination of acoustic instability, lexical diversity and temporal hesitation patterns defines degradation in motor, semantic, and executive spheres. This justifies the idea of defining speech as a systems-level biomarker and not a measure of single domain.

Second, the idea of explainability is not peripheral in clinical artificial intelligence, it is central. Smooth-working opaque systems are at the risk of being rejected in the actual healthcare setting. This study shows how computational intelligence may be consistent with clinical epistemology by integrating SHAP-based interpretability through a model pipeline, making it possible to trust, be transparent, and collaboratively make decisions with humans and AI. Third, to establish the detection of cognitive impairment early there must be a reshaping of the problem

as a trajectory modeling problem of the problem, and not a problem of different classifications.

The intermediate specificity of Mild Cognitive Impairment is an indication of the heterogeneity and subtlety of the prodromal stages. The longitudinal monitoring and individualized baselines should be included into the future frameworks to trace the progressive decline patterns.

6.3 Clinical and Societal Implications

Dementia keeps on increasing its burden on the health care systems throughout the world. One of the greatest undiscovered issues in neurology is early diagnosis. Old school methods of diagnosing depend much on neuroimaging, biomarker tests, and expert assessments, which might be expensive and unavailable in low-resource centers.

A more impressive alternative is speech-based screening. Recording speech involves minimum equipments, can be conducted remotely and is culturally flexible. The model created within this study proves that speech analysis may attain the great dementia sensitivity, so that the method could be used as the initial screening tool in the primary care or community health settings.

In addition, remote digital assessment is especially applicable to the aging population and telemedicine systems. With the transition of healthcare systems towards less invasive methods of monitoring, non-invasive digital biomarkers can be the central elements of preventive neurology.

6.4 Methodological Strengths

There are a number of methodological strengths associated with this research. The combination of acoustic, linguistic and temporal properties ensures the multimodal representation of the speech properties. The implementation of classical and deep

learning models offers fair evaluation of architectures as opposed to making use of a single paradigm. The use of the correction of class imbalance promotes fairness in the diagnostic categories. Lastly, explainability is achieved with SHAP and helps to bridge the gap between predictive analytics and clinical interpretability.

All these strengths make the proposed framework scientifically rigorous as well as translationally viable.

6.5 Future Perspectives and Research Prospect

Further studies need to be based on increasing the volume and demographics to guarantee cross-linguistic and cross-cultural generalization. Sensitivity of the early-stage of cognitive decline could greatly be enhanced with the help of longitudinal speech tracking. Also, multimodal combination with neuroimaging, genetic markers like APOE genotype, and electronic motor biomarkers may have a positive effect on predictive strength. More sophisticated modeling techniques, such as techniques of fairness-conscious learning and techniques of quantifying uncertainty, should also be considered in order to increase reliability in practice.

Finally, this study would be part of a greater paradigm shift in practice in the field of medicine: the shift toward constant clinical assessment with the help of AI and continuous monitoring of cognitive health. Through the combination of computational intelligence and neuropsychological theory and clinical ethics, this study provides the basis to the next generation of digital cognitive diagnostics.

One of the main contributions of this study is that it empirically validated the idea of cross-dataset generalization. The multimodal framework is able to shift the optimization of datasets to the clinically significant robustness by showing consistent performance on the external Pitt corpus.

Translational aspect of the role of AI diagnostics through speech supports the possibility of high ROC-AUC and F1-scores that can be obtained in independent cohorts. Such a third-party confirmation enhances the scientific integrity of the

proposed system and makes it a promising candidate of large-scale, practical applications of dementia screening in real-world scenarios.

This study introduces a generalizable multimodal AI system to detect dementia and model its progression using speech. The presence of both acoustic and linguistic information, explainable AI and cross-dataset confirmation assist in the development of reliable digital biomarkers. Future directions must be on a greater multi-centric data, longitudinal data modeling, and combine them with other complementary modalities like neuroimaging and genetic biomarkers to further improve the ability to detect the problem in its early stages.

Bibliography

[1] World Health Organization, "Dementia," WHO Fact Sheet, 2025, available at: https://www.who.int/news-room/fact-sheets/detail/dementia.

[2] I. Oiza-Zapata and A. Gallardo-Antolín, "Alzheimer's disease detection from speech using shapley additive explanations for feature selection and enhanced interpretability," *Electronics*, vol. 14, no. 11, p. 2248, 2025.

[3] K. Mekulu, F. Aqlan, and H. Yang, "Automated detection of early-stage dementia using large language models: A comparative study on narrative speech," *medRxiv*, pp. 2025–06, 2025.

[4] V. H. Jasodanand, S. S. Kowshik, S. Puducheri, M. F. Romano, L. Xu, R. Au, and V. B. Kolachalama, "Ai-driven fusion of multimodal data for alzheimer's disease biomarker assessment," *Nature Communications*, vol. 16, no. 1, p. 7407, 2025.

[5] F. García-Gutiérrez, M. Alegret, M. Marquié, N. Muñoz, G. Ortega, A. Cano, and S. Valero, "Unveiling the sound of the cognitive status: Machine learning-based speech analysis in the alzheimer's disease spectrum," *Alzheimer's Research & Therapy*, vol. 16, no. 1, p. 26, 2024.

[6] Z. Jahan, S. B. Khan, and M. Saraee, "Early dementia detection with speech analysis and machine learning techniques," *Discover Sustainability*, vol. 5, no. 1, p. 65, 2024.

[7] K. Ding, M. Chetty, A. Noori Hoshyar, T. Bhattacharya, and B. Klein, "Speech based detection of alzheimer's disease: a survey of ai techniques,

datasets and challenges," *Artificial Intelligence Review*, vol. 57, no. 12, p. 325, 2024.

[8] M. K. Vrindha, V. Geethu, P. R. Anurenjan, S. Deepak, and K. G. Sreeni, "A review of alzheimer's disease detection from spontaneous speech and text," in *2023 International Conference on Control, Communication and Computing (ICCC)*. IEEE, 2023, pp. 1–5.

[9] X. Qi, Q. Zhou, J. Dong, and W. Bao, "Noninvasive automatic detection of alzheimer's disease from spontaneous speech: a review," *Frontiers in Aging Neuroscience*, vol. 15, p. 1224723, 2023.

[10] L. Tang, Z. Zhang, F. Feng, L. Z. Yang, and H. Li, "Explainable alzheimer's disease detection using linguistic features from automatic speech recognition," *Dementia and Geriatric Cognitive Disorders*, vol. 52, no. 4, pp. 240–248, 2023.

[11] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, "Machine learning for dementia prediction: a systematic review and future research directions," *Journal of Medical Systems*, vol. 47, no. 1, p. 17, 2023.

[12] M. Parsapoor, "Ai-based assessments of speech and language impairments in dementia," *Alzheimer's & Dementia*, vol. 19, no. 10, pp. 4675–4687, 2023.

[13] Z. Shah, S. A. Qi, F. Wang, M. Farrokh, M. Tasnim, E. Stroulia, and A. Katsamanis, "Exploring language-agnostic speech representations using domain knowledge for detecting alzheimer's dementia," in *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–2.

[14] J. Robin, M. Xu, A. Balagopalan, J. Novikova, L. Kahn, A. Oday, and E. Teng, "Automated detection of progressive speech changes in early alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 15, no. 2, p. e12445, 2023.

[15] U. Petti, R. Nyrup, J. M. Skopek, and A. Korhonen, "Ethical considerations in the early detection of alzheimer's disease using speech and ai," in

Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. ACM, 2023, pp. 1062–1075.

[16] Q. Yang, X. Li, X. Ding, F. Xu, and Z. Ling, "Deep learning-based speech analysis for alzheimer's disease detection: a literature review," *Alzheimer's Research & Therapy*, vol. 14, no. 1, p. 186, 2022.

[17] Tahouramorovati, "Dementia detection using speech dataset," Kaggle dataset, 2024, available at: https://www.kaggle.com/datasets/tahouramorovati/dementia-detection-using-speech.

[18] "Dementiabank pitt corpus," TalkBank dataset, DementiaBank English Pitt Corpus, 1994, available at: https://talkbank.org/dementia/access/English/Pitt.html (accessed Month Day, Year). Supported by NIH grants AG03705 and AG05133.

[19] M. R. Kumar, S. Vekkot, S. Lalitha, D. Gupta, V. J. Govindraj, K. Shaukat, and M. Zakariah, "Dementia detection from speech using machine learning and deep learning architectures," *Sensors*, vol. 22, no. 23, p. 9311, 2022.

[20] A. Ševčík and M. Rusko, "A systematic review of alzheimer's disease detection based on speech and natural language processing," in *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2022, pp. 01–05.

[21] A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Exploring dementia detection from speech: Cross corpus analysis," in *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 6472–6476.

[22] F. Bertini, D. Allevi, G. Lutero, L. Calzà, and D. Montesi, "An automatic alzheimer's disease classifier based on spontaneous spoken english," *Computer Speech & Language*, vol. 72, p. 101298, 2022.

[23] F. Agbavor and H. Liang, "Predicting dementia from spontaneous speech using large language models," *PLOS Digital Health*, vol. 1, no. 12, 2022.

[24] A. Favaro, S. Motley, Q. M. Samus, A. Butala, N. Dehak, E. S. Oh, and L. Moro-Velazquez, "Artificial intelligence tools to evaluate language and speech patterns in alzheimer's disease," *Alzheimer's & Dementia*, vol. 18, p. e064913, 2022.

[25] R. Li, X. Wang, K. Lawler, S. Garg, Q. Bai, and J. Alty, "Applications of artificial intelligence to aid early detection of dementia: a scoping review on current capabilities and future directions," *Journal of Biomedical Informatics*, vol. 127, p. 104030, 2022.

[26] E. Fristed, C. Skirrow, M. Meszaros, R. Lenain, U. Meepegama, S. Cappa, and J. Weston, "A remote speech-based ai system to screen for early alzheimer's disease via smartphones," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 14, no. 1, p. e12366, 2022.

[27] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153–4164, 2022.

[28] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech," *Frontiers in Computer Science*, vol. 3, p. 780169, 2021.

[29] Z. Liu, Z. Guo, Z. Ling, and Y. Li, "Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation," in *ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 7323–7327.

[30] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 670–677.

[31] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 642647, 2021.

[32] M. Antonsson, K. Lundholm Fors, M. Eckerström, and D. Kokkinakis, "Using a discourse task to explore semantic ability in persons with cognitive impairment," *Frontiers in Aging Neuroscience*, vol. 12, p. 607449, 2021.

[33] L. Calzà, G. Gagliardi, R. R. Favretti, and F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, p. 101113, 2021.

[34] C. Xue, C. Karjadi, I. C. Paschalidis, R. Au, and V. B. Kolachalama, "Detection of dementia on voice recordings using deep learning: a framingham heart study," *Alzheimer's Research & Therapy*, vol. 13, no. 1, p. 146, 2021.

[35] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Zelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios," in *Interspeech 2021*. ISCA, 2021, pp. 3825–3829.

[36] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection," in *Interspeech 2021*. ISCA, 2021, p. 3790.

[37] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 623607, 2021.

[38] S. El-Sappagh, J. M. Alonso, S. R. Islam, and Sultan, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease," *Scientific Reports*, vol. 11, p. 2660, 2021.

[39] L. V. Srinivas, R. S. Shankar, S. Rajeswari, and et. al., "Enhancing dementia detection (edd): A machine learning ensemble approach for early alzheimer's diagnosis," in *Advances in Electrical and Computer Technologies*. CRC Press, 2026, pp. 96–105.

[40] M. S. S. Syed, Z. S. Syed, and et al., "Automated screening for alzheimer's dementia through spontaneous speech," in *Interspeech 2020*. ISCA, 2020.

[41] H. Shao, Y. Pan, Y. Wang, and Y. Zhang, "Modality fusion using auxiliary tasks for dementia detection," *Computer Speech & Language*, vol. 95, p. 101814, 2026.

[42] A. H. Abdulaal, M. Valizadeh, and M. C. Amirani, "Neurophysiological biomarker extraction through decoupled attention mechanisms and eeg signal processing in alzheimer's disease and frontotemporal dementia," *Biomedical Signal Processing and Control*, vol. 117, p. 109559, 2026.

[43] A. H. Abdulaal, Valizadeh, and et al, "Predictive models for early detection and prognosis of dementia using artificial intelligence and machine learning," in *Next-Gen Healthcare: AI-Powered Medical Innovations*. Cham, Switzerland: Springer Nature, 2026, pp. 219–237.

[44] P. Govindarajan, S. K. KT, A. U. Menon, A. K. Das, A. KR, and A. Tesfahun, "Dementia detection from speech using cogni-wave," in *Proceedings of the 12th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2025, pp. 1–9.

[45] I. Cohen, R. A. Taylor, H. Xue, I. V. Faustino, N. Festa, C. Brandt, E. Gao, L. Han, S. Khasnavis, J. M. Lai *et al.*, "Detection of emergency department patients at risk of dementia through artificial intelligence," *Alzheimer's & Dementia*, vol. 21, no. 6, p. e70334, 2025.

[46] T. A. Collyer, M. Liu, R. Beare, Andrew *et al.*, "Dual-stream algorithms for dementia detection: Harnessing structured and unstructured electronic health record data, a novel approach to prevalence estimation," *Alzheimer's & Dementia*, vol. 21, no. 5, p. e70132, 2025.

[47] A. C. Mmadumbu, F. Saeed, F. Ghaleb, and S. N. Qasem, "Early detection of alzheimer's disease using deep learning methods," *Alzheimer's & Dementia*, vol. 21, no. 5, p. e70175, 2025.

[48] S. Naole, D. Parikh, S. Nayak, and S. P. Ramu, "Evaluating cognitive assessment tools: A comparative analysis of mmse, rudas, sage, adas and moca for early dementia detection," *arXiv preprint*, p. arXiv:2505.07246, 2025.

[49] M. Norouzi, R. Kafieh, P. Chazot, D. T. Smith, and Z. Amini, "Insights from the eyes: A systematic review and meta-analysis of the intersection between eye-tracking and artificial intelligence in dementia," *Aging & Mental Health*, vol. 29, no. 8, pp. 1367–1375, 2025.

[50] V. K. Choudhary, A. Pawar, S. Suman, G. Kumar, A. Kumar, and P. Kumar, "A comprehensive review on artificial intelligence for human brain disease," in *Proceedings of the International Conference on Inventive Computation Technologies (ICICT)*, 2025, pp. 690–696.

[51] V. M. Joshi, P. P. Dandavate, R. Rashmi, G. R. Shinde, D. D. Kulkarni, and R. Mirajkar, "Demnet neurodeep: Alzheimer detection using electroencephalogram and deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 457–465, 2025.

[52] R. Hafeez, S. Waheed, S. A. Naqvi, F. Maqbool, A. Sarwar, S. Saleem, M. I. Sharif, K. Siddique, and Z. Akhtar, "Deep learning in early alzheimer's disease's detection: A comprehensive survey of classification, segmentation, and feature extraction methods," *arXiv preprint arXiv:2501.15293*, 2025.

[53] J. Wang, M. Choi, P. Buto, J. D. Kelly, R. La Joie, J. Kornak, S. C. Zimmerman, R. Chen, E. Raphael, C. A. Schaefer *et al.*, "Detection bias in ehr-based research on clinical exposures and dementia," *JAMA network open*, vol. 8, no. 4, p. e256637, 2025.

[54] W. Hechkel and A. Helali, "Early detection and classification of alzheimer's disease through data fusion of mri and dti images using the yolov11 neural network," *Frontiers in Neuroscience*, vol. 19, p. 1554015, 2025.

[55] I. Nagarajan and G. G. Lakshmi Priya, "A comprehensive review on early detection of alzheimer's disease using various deep learning techniques," *Frontiers in Computer Science*, vol. 6, p. 1404494, 2025.

[56] S. Mohsen, "Alzheimer's disease detection using deep learning and machine learning: a review," *Artificial Intelligence Review*, vol. 58, no. 9, p. 262, 2025.

[57] C. Gettel, J. Galske, K. Araujo, S. Dresden, J. Dussetschleger, L. Iannone, J. Lai, P. Martin, B. Mignosa, K. Muschong *et al.*, "Detection and differentiation of undiagnosed dementia in the emergency department: a pilot referral pathway," *Alzheimer's & Dementia*, vol. 21, no. 4, p. e70189, 2025.

[58] K. Stefanou, K. D. Tzimourta, C. Bellos, G. Stergios, K. Markoglou, E. Gionanidis, M. G. Tsipouras, N. Giannakeas, A. T. Tzallas, and A. Miltiadous, "A novel cnn-based framework for alzheimer's disease detection using eeg spectrogram representations," *Journal of Personalized Medicine*, vol. 15, no. 1, p. 27, 2025.

[59] N. R. Fowler, K. A. Partrick, J. Taylor, M. Hornbecker, K. Kelleher, M. Boustani, J. L. Cummings, T. MacLeod, M. M. Mielke, J. R. Brosch *et al.*, "Implementing early detection of cognitive impairment in primary care to improve care for older adults," *Journal of internal medicine*, vol. 298, no. 1, pp. 31–45, 2025.

[60] K. R. Krishnan, R. M. Levy, H. R. Wagner, G. Chen, K. Gersing, and P. M. Doraiswamy, "Informant-rated cognitive symptoms in normal aging, mild cognitive impairment, and dementia: Initial development of an informant-rated screen (brief cognitive scale) for mild cognitive impairment and dementia," *Psychopharmacology Bulletin*, vol. 35, no. 3, pp. 79–88, 2001.

[61] R. M. Wahul, S. Ambadekar, D. M. Dhanvijay, Dhanvijay *et al.*, "Multimodal approaches and ai-driven innovations in dementia diagnosis: a systematic review," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 96, 2025.

[62] F. Akbar, I. Taj, S. M. Usman *et al.*, "Unlocking the potential of eeg in alzheimer's disease research: Current status and pathways to precision detection," *Brain Research Bulletin*, vol. 223, p. 111281, 2025.

[63] K. A. Cody, L. Du, R. L. Studer, E. M. Jonaitis, S. Asthana, B. T. Christian, N. A. Chin, K. M. Kirmess, M. R. Meyer, K. E. Yarasheski *et al.*, "Accuracy of plasma biomarkers to detect alzheimer's disease proteinopathy prior to dementia," *Alzheimer's & Dementia*, vol. 21, no. 3, p. e14570, 2025.