Leveraging Multimodal Machine Learning on Standardized Data for Early Dementia Detection

Guided Research

Chair of Software Engineering for Business Information Systems
Supervisor: Alexandre Mercier
Examiner: Prof. Dr. Florian Matthes

Marcel Seitz
marcely.seitz@tum.de

Abstract

Early detection of Mild Cognitive Impairment (MCI) is essential for timely intervention before progression to dementia. This study presents automatic MCI detection based on the Delaware corpus within DementiaBank, focusing on the Cookie Theft Picture Description task. A reproducible multimodal framework was developed to align and process 1,386 paired audio–text segments, representing audio as log-Mel spectrograms or waveforms and text as either orthographic or phonetic (ARPAbet) transcripts. Unimodal audio (ViT, CNN, HuBERT), unimodal text (BERT, DistilBERT), and multimodal cross-attention fusion configurations were systematically compared. Hyperparameters were optimized via Bayesian optimization, and statistical validation was performed using Welch's unequal-variance t-tests on participant-level splits. The Vision Transformer (ViT) achieved the highest F1 (0.77) and balanced accuracy (0.65), confirming the strong discriminative power of acoustic features. Phonetic ARPAbet transcriptions improved BERT stability, particularly within multimodal fusion frameworks, while DistilBERT performed best on orthographic text. Among multimodal models, CNN–BERT (non-ARPA) achieved the highest balanced accuracy (0.65) and precision (0.64), showing small but statistically significant differences ($p < 0.001$) relative to the best unimodal model. Acoustic explanations revealed interpretable spectro–temporal patterns consistent with clinical markers of cognitive decline, underscoring the potential of speech-based screening for early detection of cognitive impairment.

Keywords: mild cognitive impairment, DementiaBank, Cookie Theft Picture Description, multimodal fusion, cross-attention, Vision Transformer, BERT, phonetic transcription (ARPAbet)

1 Introduction

Early detection of cognitive decline is a major objective in neurodegenerative research, as timely identification of Mild Cognitive Impairment (MCI) can enable preventive interventions before progression to dementia [42]. Traditional diagnostic methods rely on neuroimaging and neuropsychological testing, whereas speech provides a non-invasive, cost-efficient, and easily deployable alternative for monitoring cognitive status [2, 10, 33, 39, 46]. Subtle changes in lexical diversity, syntactic complexity, fluency, and prosody have been shown to indicate early cognitive deterioration [2, 10, 32, 33, 39, 46, 48, 49]. Building on these findings, recent community challenges such as the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) and its follow-up ADReSSo have introduced standardized datasets and evaluation protocols for automatic Alzheimer's detection and MMSE regression from spontaneous speech [27, 28]. However, these initiatives primarily target Alzheimer's dementia rather than the prodromal MCI stage and typically model either acoustic or linguistic features in isolation [2, 10, 32, 33, 39, 45, 46]. Systematic investigations of integrated audio–text modeling for distinguishing cognitively healthy controls (HC) from individuals with MCI remain limited.

To address this gap, this study introduces a unified multimodal framework for classifying MCI versus HC from the Cookie Theft Picture Description (CTP) task [9] within the Delaware corpus of DementiaBank [24]. The framework temporally aligns and processes paired audio–text segments and enables systematic comparison across unimodal and multimodal configurations. Three audio encoders (Vision Transformers, CNN, HuBERT) and two text encoders (BERT, DistilBERT) are implemented under both orthographic and phonetic (ARPAbet) transcription conditions to evaluate complementary acoustic–linguistic cues. Hyperparameters are optimized via Bayesian optimization using Optuna [1], and performance is statistically compared using Welch's unequal-variance t-tests [52] across trials.

Accordingly, the study investigates (1) the discriminative power of unimodal audio and text models, (2) the benefit of multimodal cross-attention fusion, and (3) the impact of orthographic versus phonetic transcriptions on model performance and stability.

2 Theoretical Background

2.1 Clinical Background of Cognitive Impairment

MCI represents an intermediate state between healthy aging and dementia, characterized by measurable decline in one or more cognitive domains while functional independence is preserved [2, 12, 39, 42, 48]. Longitudinal studies indicate that 19–50% of individuals with MCI progress to Alzheimer's disease within three to five years [8, 16, 30], highlighting the importance of early detection. However, MCI outcomes are heterogeneous, with some individuals remaining stable or reverting to normal cognition [22, 31]. While Alzheimer's dementia involves pronounced anomia and discourse deficits, MCI manifests subtler lexical retrieval and semantic impairments that often precede overt decline [22, 31]. Speech and language are therefore promising non-invasive markers of prodromal cognitive change [2, 10, 32, 33, 46].

2.2 Speech and Language as early Biomarkers of Cognitive Decline

Speech production involves multiple cognitive processes such as memory, attention, executive control, and linguistic formulation, making it highly sensitive to neurocognitive deterioration [2, 10, 32, 48, 49, 55]. In MCI, studies consistently report reduced lexical diversity, reliance on high-frequency or vague terms, and word-finding difficulties [9, 10, 32, 46, 48]. Utterances become shorter and syntactically simpler, and fluency decreases due to hesitations, pauses, and self-corrections [2, 32, 48, 49]. Acoustic changes such as slower speech rate, longer pauses, flatter intonation, and reduced articulatory precision are also observed [10, 33]. Together, these linguistic and prosodic features form measurable indicators of early cognitive decline.

2.3 Cookie Theft Picture Description Task

The CTP task from the Boston Diagnostic Aphasia Examination [9] is a standard method for eliciting connected speech under controlled but naturalistic conditions [4, 32, 46, 48]. By describing a fixed visual scene, speakers reveal semantic, syntactic, and pragmatic patterns that differ across cognitive conditions [2, 32, 46]. The fixed stimulus minimizes memory and attentional demands, enabling reliable comparison of discourse organization and informativeness between healthy and cognitively impaired groups [32, 48].

2.4 Machine Learning Approaches for Cognitive Impairment Detection

Machine learning techniques have been widely applied to speech-based cognitive assessment [2, 28, 33, 39, 46]. Acoustic approaches often use Mel-frequency cepstral coefficients, prosodic descriptors such as pitch, energy, and pause statistics, and standardized feature sets like eGeMAPS to capture voice markers of decline [39, 46, 49]. Text-based models leverage pretrained Transformers such as BERT and DistilBERT to model contextual semantics and mitigate data sparsity [39, 45]. Recent multimodal studies combine audio and text using feature concatenation, ensembling, or attention-based fusion to exploit complementary information [2, 33, 46]. These developments position ML as a scalable and non-invasive complement to traditional clinical diagnostics.

2.5 Challenges in Multimodal Dementia Detection

Despite recent progress, multimodal dementia detection faces persistent challenges. Available datasets are small and imbalanced, increasing the risk of overfitting [2, 12, 33]. Participant-level splits are often neglected, which can lead to speaker leakage [33]. Precise temporal alignment between modalities remains difficult, and inconsistent evaluation practices hinder comparability [2, 32, 46]. Finally, most frameworks lack integrated explainability that links model attributions to clinically interpretable speech markers [10].

3 Methodology

3.1 Dataset

The dataset analyzed in this study was derived from the Delaware corpus within DementiaBank, a clinically curated repository comprising discourse elicitation tasks recorded from individuals diagnosed with MCI and HC [24]. All recordings and transcripts are in English. Participants range in age from approximately 60 to 90 years, with a predominance of female speakers. Cognitive status was

assessed in the original corpus using the Montreal Cognitive Assessment, with scores between 19 and 30 (clinical cutoff = 26), confirming that the original cohort included only HC and MCI participants. The analyzed subset includes 64 MCI and 48 HC participants, restricted to samples corresponding to the CTP task. Only participant utterances were included, while interviewer turns and non-task discourse were excluded.

Each CTP recording is accompanied by a manual transcript in the CHAT format [29], containing linguistic content, time annotations, and participant metadata. Using these timecodes, utterance-level segments were extracted to create synchronized audio–text pairs annotated with diagnostic labels. The resulting dataset comprised 1,386 paired segments with an average duration of 3.17 s (maximum 19.10 s), representing multiple temporally ordered utterances per participant.

To prevent participant leakage and ensure unbiased evaluation, the data were split at the participant level into disjoint training, validation, and test partitions (70%, 15%, 15%). Stratified sampling on diagnostic labels (MCI vs. HC) was used to maintain class balance across subsets, and multiple deterministic random seeds were tested until the test partition achieved a minority-to-majority ratio above 0.9. The final splits were approximately balanced, with the segment-level test set comprising 149 HC and 138 MCI samples.

3.2 Data Preprocessing

The preprocessing workflow (Figure 1) transforms raw audio and transcript data from the Delaware corpus into synchronized, model-ready inputs for unimodal and multimodal modeling. It comprises standardized audio processing, text normalization, tokenization, augmentation, phonetic alignment, and temporal synchronization across modalities.

3.2.1 Audio Preprocessing

Standardization. All recordings were resampled to 16 kHz and converted to single-channel waveforms for consistency across participants. Segment boundaries were derived from transcript timecodes to ensure precise temporal alignment between audio and text. To ensure consistent input length during training, audio segments longer than the 95^{th}-percentile length (≈ 8.3 s) were truncated, and shorter ones were zero-padded.

Feature Extraction. Log-Mel spectrograms were computed using 128 Mel filters, a 2 048-sample Hann window, 512 -sample hop length, and 2 048-sample-point fast Fourier transform, providing balanced time–frequency resolution at 16 kHz [18]. The resulting spectrograms were converted to the decibel (log) scale and later resized to 224×224 pixels during data loading to match the input resolution of convolutional and transformer backbones. Each spectrogram constituted a standardized single-channel feature representation for model input.

Data Augmentation. Offline spectrogram augmentations included light frequency masking, adapted from the SpecAugment framework [35], while omitting time masking and warping to preserve alignment with transcript boundaries and maintain lexical–acoustic correspondence across modalities. This procedure perturbed only spectral energy distributions while preserving the original segment duration and spectrogram shape. Both original and augmented log-Mel spectrograms were stored for subsequent training use (see Section 3.4).

3.2.2 Text Preprocessing

Transcript Representations. Two complementary textual forms were prepared: (1) the original orthographic transcripts and (2) phoneticized ARPAbet counterparts. Orthographic transcripts, extracted from the CHAT-formatted files, were cleaned by removing nonlinguistic annotations, filler tags, and special symbols. Disfluencies such as *uh* and *um* were retained, as they constitute clinically informative markers of fluency disruption and lexical retrieval difficulty in MCI speech [49]. Text was lowercased, punctuation-normalized, and segments containing fewer than twenty characters were excluded. For brevity, orthographic inputs are referred to as *non-ARPA* in the result tables (see Section 4)

Tokenization. Tokenization was performed dynamically during data loading to ensure consistent truncation and padding before model input, which conceptually belongs to the preprocessing. Text sequences were tokenized using pretrained transformer tokenizers [54] with a fixed sequence length of 128 tokens, following common sentence-level classification practice. In unimodal text experiments, complete transcript segments were tokenized, whereas in multimodal setups, orthographic and ARPAbet strings were truncated or padded based on pre-aligned temporal segments (see Sections 3.2.3–3.2.4).

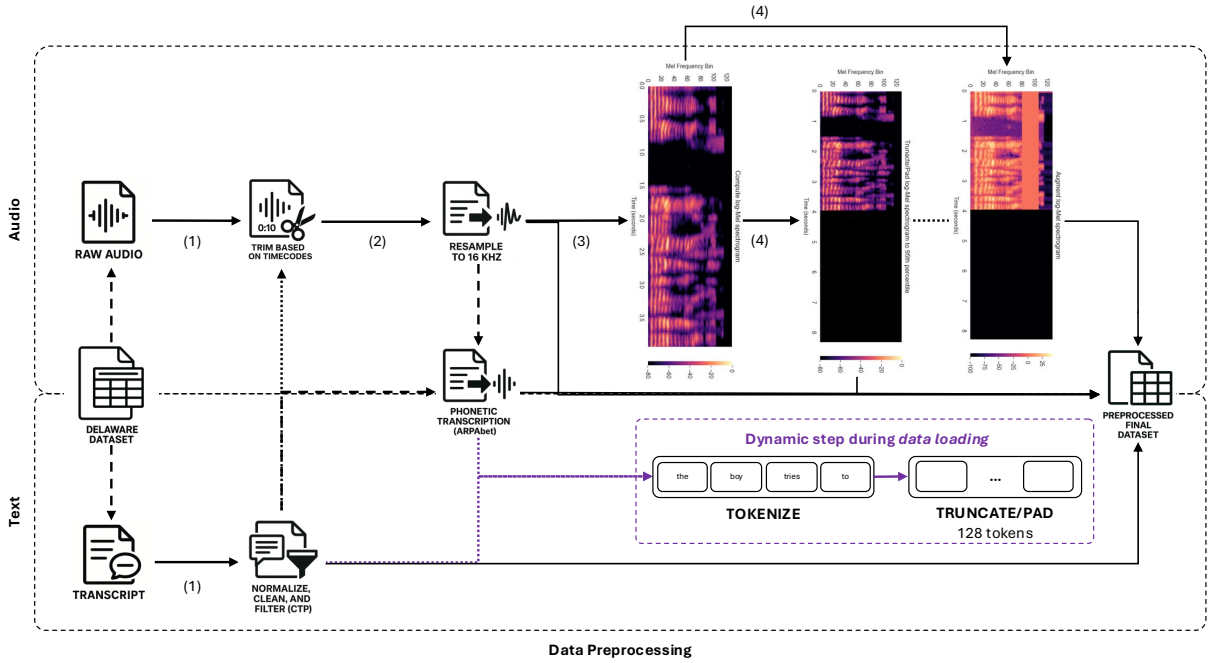

Figure 1: Overview of the data preprocessing pipeline. Alignment details are omitted for clarity (see Section 3.2.4).

3.2.3 Phonetic Preprocessing

Forced Alignment. Phonetic representations were generated using the Gentle forced aligner [34], executed in a controlled environment to ensure cross-platform reproducibility. For each participant, Gentle aligned the cleaned transcript text segments to the corresponding audio at word and phoneme levels, producing time-stamped boundaries. The resulting phoneme sequences were converted to ARPAbet strings, which served as phoneticized text for all ARPA-based experiments. For brevity, these phoneticized inputs are referred to as *ARPA* in the result tables (see Section 4).

3.2.4 Temporal Alignment.

For multimodal experiments, both orthographic and phonetic transcripts were temporally aligned with their corresponding audio segments using transcript-based timecodes at the word level to ensure lexical–acoustic synchronization. When a boundary intersected a spoken word, the entire word was retained to preserve semantic and phonetic integrity. Text spans were proportionally truncated to match the normalized audio duration (as in Section 3.2.1). This adjustment was applied only in multimodal setups; unimodal text models used the full-length transcript segments.

3.3 Model Architectures

As illustrated in Figure 2, the developed framework comprises three processing branches: an audio branch, a text branch, and a multimodal cross-attention fusion branch. All model architectures were implemented in PyTorch [37]. Backbone choices and architectural design were guided by prior research on DementiaBank and related pathological speech datasets [2]. Depending on the configuration, the output layer produced either two unnormalized logits (for softmax-based classifiers) or a single scalar logit (for sigmoid-based classifiers), representing the probability of the MCI class. Details of optimization and loss formulations are provided in Section 3.4.

3.3.1 Unimodal Audio

Audio encoders were implemented using three backbone families: Vision Transformers (ViT), convolutional neural networks (CNNs), and HuBERT. Both ViT and CNN models operated on log-Mel spectrogram representations (224×224), capturing time–frequency patterns through complementary architectural principles: global self-attention for ViT and local convolutional filtering for CNNs. ViT models followed the standard implementation provided in the `timm` library [13, 53] with 16×16 patch embeddings, where the small variant served as the default backbone and tiny/base versions were explored during optimization (see Section 3.5).

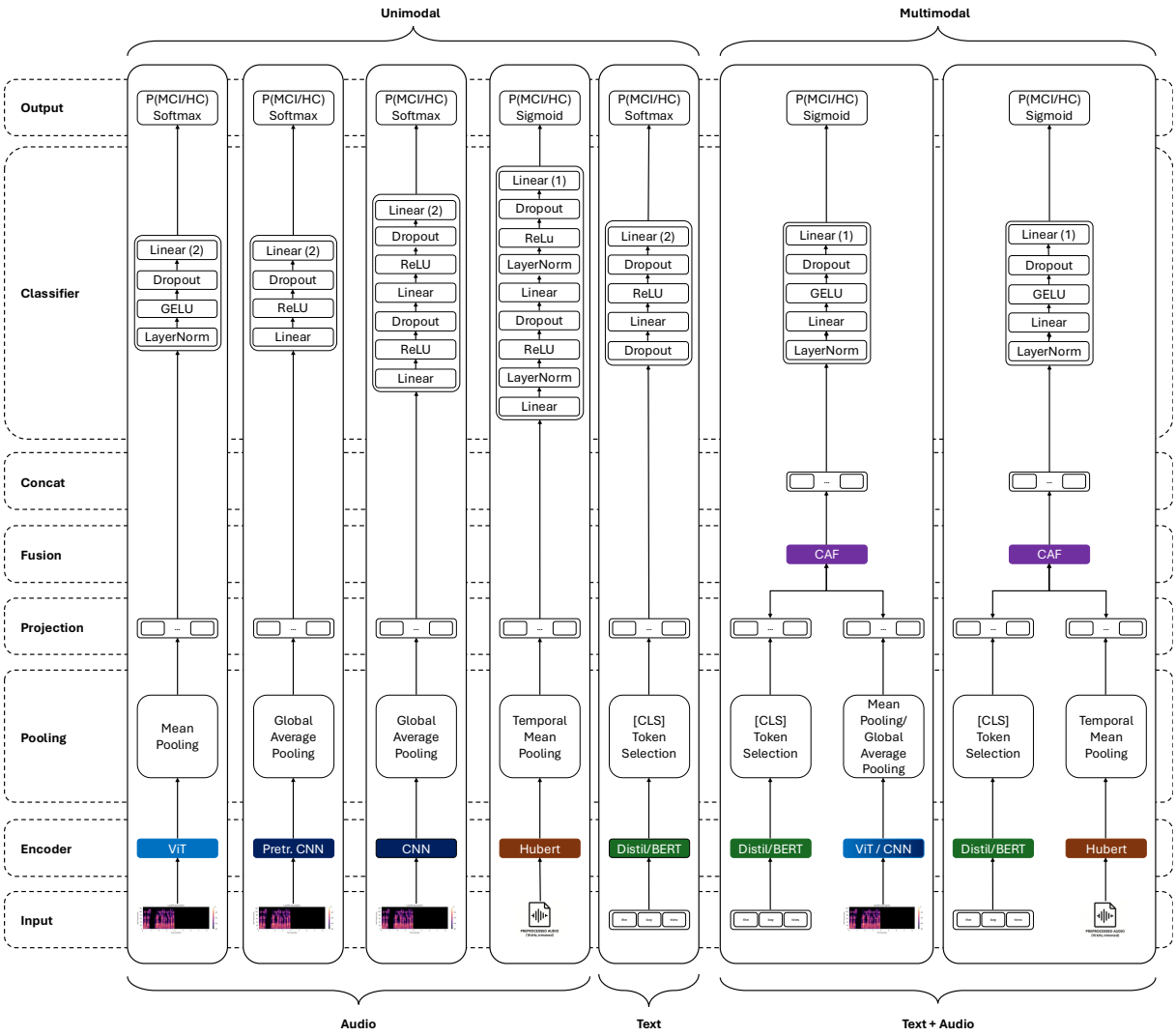

Figure 2: Overview of the unimodal and multimodal architectures. Each branch corresponds to an encoder configuration (audio, text, or joint audio–text fusion). For brevity, BERT and DistilBERT are shown as a shared text branch.

CNN encoders included a custom four-block CNN and a ResNet-18 variant adapted for single-channel input [19]. HuBERT (base) operated directly on raw 16 kHz waveforms [21], generating transformer-based acoustic representations with 768 hidden units, with lower layers optionally frozen for stability (see Section 3.4). In unimodal experiments, each audio encoder acted as a standalone end-to-end classifier trained on its respective input representation.

3.3.2 Unimodal Text

Text encoders were based on pretrained BERT-base-uncased and DistilBERT-base-uncased models [11, 44]. The uncased variants were chosen to reduce sparsity and improve robustness to capitalization differences in spontaneous speech transcripts. The `[CLS]` token from the final transformer layer was used as a sentence-level representation and passed through a two-layer classification head (64-unit ReLU, dropout, and linear output). The transformer backbone could be either frozen or fine-tuned depending on the experimental configuration (see Section 3.4).

3.3.3 Multimodal Audio–Text Fusion

For multimodal modeling, the same audio and text encoders were reused as feature extractors. Their outputs were projected into a shared 256-dimensional latent space and fused via a Cross-Attention Fusion (CAF) module [50, 51], allowing each modality to attend to salient features of the other. The attended representations were concatenated (512 dimensions), normalized, and processed by a 128-unit GELU layer [20] with dropout before the final linear classifier. This design allows

flexible substitution of the different audio and text backbones without modifying the overall training pipeline.

Design Rationale. Given the limited size and heterogeneity of the dataset, the CAF module was implemented as a single-head bidirectional attention mechanism with residual connections to ensure stable gradient propagation, rather than a deeper multi-head transformer. This lightweight design was adopted to reduce overfitting risk and improve interpretability while maintaining stable convergence on small clinical datasets. This choice aligns with prior findings that shallow cross-modal attention is sufficient for effective fusion under low-data conditions (cf. [6, 50]). Moreover, since both the audio and text encoders are pretrained transformer architectures that already employ multi-layer self-attention internally [11, 13, 21, 44], additional multi-head fusion was considered redundant given their existing representational capacity.

3.4 Training Setup

Optimization. All models were optimized using the AdamW optimizer, chosen for its stable convergence and effective decoupling of weight decay from the adaptive learning rate [26]. Optimization hyperparameters (learning rate, batch size, dropout, weight decay) were defined via configuration settings and, when applicable, tuned through Bayesian optimization [1] (see Section 3.5 for details). A dynamic learning-rate scheduler reduced the learning rate when the validation objective (i.e., the composite score described below) plateaued, and early stopping terminated training after a predefined number of non-improving epochs [40]. Together, these mechanisms enhanced stability and prevented overfitting under limited data conditions. For text and multimodal experiments, a configuration option specified whether models were trained on cleaned orthographic transcripts or ARPAbet phonetic transcriptions. Optional encoder freezing policies for HuBERT and text backbones were controlled via configuration flags.

Training Augmentation. Precomputed spectrogram augmentations were sampled during training to improve robustness and mitigate overfitting. For spectrogram-based models (CNN, ViT), roughly half of training items were replaced by their precomputed frequency-masked counterparts. For waveform-based models (HuBERT), online augmentations (Gaussian noise, pitch shift, temporal shift, and gain variation) were applied stochastically per sample, each with an independent activation probability. All augmentations preserved segment duration to maintain alignment. Validation and test sets were evaluated on unaltered signals.

Threshold Optimization and Model Selection. Validation performance was monitored using a composite score combining binary F1, balanced accuracy, and precision (weights 0.4, 0.4, 0.2). At each epoch, the optimal decision threshold was determined by sweeping a predefined range on the validation set and maximizing this composite objective [43]. The resulting epoch-specific threshold was then applied to compute all validation metrics, ensuring consistent evaluation at the optimal operating point. The combined use of binary F1 and balanced accuracy ensured stable model selection by jointly capturing precision–recall balance and class-level sensitivity under imbalance (cf. [43]). Although precision contributes to the F1 metric, it was included separately to offset recall-oriented configurations and preserve clinically relevant specificity. The best-performing model according to this composite validation score was retained for subsequent analyses.

Loss Functions. Loss functions were modality-specific and class-imbalance aware. Unimodal spectrogram-based and text models used weighted cross-entropy with class weights computed via the scikit-learn [38] `balanced` scheme on the training split, and produced two logits. HuBERT-based and multimodal models produced a single scalar logit and used binary cross-entropy with logits, with the positive-class weight set to the ratio of negatives to positives in the training split. Both formulations balanced gradient contributions under moderate label imbalance.

Experiment Tracking. All experiments were logged (offline) with Weights & Biases [7]. Per-epoch metrics (loss, accuracy, precision, recall, F1, and balanced accuracy) were computed using the `scikit-learn` [38] metrics API, while optimal decision thresholds and learning-rate schedules were determined within the PyTorch training loop. Validation confusion matrices were exported each epoch, and all configurations, model checkpoints (best/final), and logs were versioned and timestamped.

3.5 Hyperparameter Optimization

Optimization Strategy. Hyperparameter optimization combined manual range exploration with automated Bayesian search using Optuna [1]. Optuna maximized the validation F1-score, while early stopping and model checkpointing during training were governed by a composite validation objective combining binary F1, balanced accuracy, and precision (weights 0.4, 0.4, 0.2; see Section 3.4). This distinction mitigated overfitting to a single metric and promoted balanced sensitivity and specificity.

Parameter Space Definition. Stable search ranges for key hyperparameters (learning rate, dropout rate, batch size, weight decay, and early-stopping patience) were empirically established through preliminary experiments. These intervals served as bounded search spaces across audio (ViT, CNN, HuBERT) and text (BERT, DistilBERT) encoders. Model-specific categorical switches, such as pretrained initialization or encoder freezing, were included when applicable.

Unimodal Optimization. Unimodal optimization was conducted in two sequential phases, controlled via configuration flags: an initial architecture search and subsequent fine-tuning. Architecture (Arch) search was applied primarily to ViT-based models, treating the backbone variant (`tiny`, `small`, `base`) as a categorical parameter while jointly optimizing continuous hyperparameters within broad predefined ranges. The best-performing ViT variant from this stage defined the architecture used in the fine-tuning phase. Fine-tuning then refined the parameter space by narrowing each numeric hyperparameter to the empirically most promising interval observed across the top 10 trials of the architecture search. During fine-tuning, Optuna jointly optimized learning rate, dropout rate, batch size, weight decay, and early-stopping patience, while categorical switches tested relevant options such as pretrained initialization (ViT, CNN) or feature-extractor freezing (HuBERT). Text models (BERT-base and DistilBERT-base) did not undergo architecture search but were fine-tuned separately under orthographic and ARPAbet input conditions using analogous hyperparameter and encoder-freezing configurations.

Multimodal Optimization. Multimodal optimization followed the same Optuna-based fine-tuning procedure as the unimodal experiments. The audio branch (ViT, CNN, or HuBERT) was optimized within a shared search space, while the text encoder (BERT or DistilBERT) remained fixed per configuration. Each multimodal configuration was evaluated for both orthographic and ARPAbet transcriptions to assess the influence of phonetic representations on cross-modal learning. Categorical switches controlled encoder freezing, and for ViT-based runs, the backbone variant could be optionally sampled (`tiny`, `small`, `base`). The cross-attention fusion module and projection dimensionality (256) remained fixed to ensure comparability across backbones.

Experiment Execution and Logging. Each Optuna study executed 100 trials per configuration, with all trial metrics and parameter settings automatically logged for evaluation and subsequent statistical analysis.

3.6 Reproducibility

All preprocessing, training, and optimization steps were executed deterministically to ensure full experimental reproducibility. Global random seeds were fixed across all components to guarantee deterministic behavior in data loading, augmentation, and model initialization. Participant-level splits were strictly enforced to prevent data leakage between subsets, and all configuration settings, model checkpoints, and log artifacts were versioned and timestamped to ensure complete traceability.

3.7 Post-Hoc Evaluation

After hyperparameter optimization, all Optuna study outputs were subjected to a standardized post-processing procedure designed to assess hyperparameter stability and to statistically compare model families and input representations. Analyses were conducted separately for unimodal audio, unimodal text, and multimodal experiments.[1]

Based on these analyses, the best-performing configurations were retrained on the combined training and validation data and subsequently evaluated once on the held-out test set to verify generalization.

Unimodal Audio Evaluation. For the unimodal audio experiments, Optuna trials were aggregated across the three backbone families (ViT, CNN, and HuBERT). Summary statistics and kernel density estimates (KDEs) [36] were computed for each performance metric to characterize distributional ten-

[1] Complete evaluation outputs, including summary tables and visualizations (e.g., kernel-density, violin, swarm, and correlation plots), are available upon request.

dencies across 100 trials per configuration. Parameter–metric relationships (e.g., learning rate, dropout rate, pretrained status) were examined to identify potential sensitivity trends, which were retained only for descriptive reporting. Pairwise performance differences between backbones were additionally evaluated using Welch's unequal-variance t-tests [52] to assess statistical significance.

Unimodal Text Evaluation. For the unimodal text experiments, Optuna trials were analyzed analogously to the unimodal audio evaluation, grouping results by transformer backbone (BERT-base vs. DistilBERT) and text representation (orthographic vs. ARPAbet). Welch's unequal-variance t-tests were pre-specified to test for potential effects of phonetic versus orthographic representations across all validation metrics.

Multimodal Evaluation. For the multimodal experiments, analyses encompassed all audio–text backbone combinations (ViT, CNN, HuBERT with BERT or DistilBERT) under both text modalities (orthographic and ARPAbet). Pairwise Welch's unequal-variance t-tests were applied to evaluate the influence of text representation within each audio architecture. Validation metrics were summarized per configuration, and cross-modal hyperparameter patterns (e.g., dropout, learning rate) were analyzed to delineate stable regions within the multimodal search space.

Consistency Across Analyses. All procedures were predefined and applied consistently across unimodal and multimodal branches to ensure comparable statistical treatment of Optuna outcomes. Pairwise performance differences were evaluated using Welch's unequal-variance t-tests, which were selected for their robustness to unequal variances. Given the extensive number of model configurations, exhaustive factorial testing was impractical; hence, Welch's tests were employed as exploratory analyses to identify consistent performance trends across configurations.

3.8 Hardware

Experiments were executed on both CPU and GPU environments. Local preprocessing and lightweight evaluations were conducted on an Apple M4 Mini with 16 GB of unified memory, whereas all deep learning experiments were trained on an NVIDIA Tesla V100 GPU (16 GB VRAM, CUDA 12.6). High-memory GPU resources were required primarily for multimodal configurations due to the increased batch and feature tensor dimensions.

4 Results

This section reports the quantitative performance of all unimodal and multimodal models, including validation metrics and pairwise significance tests based on Welch's unequal-variance t-test. Unless otherwise stated, bold values denote the best result per metric within each table, and bold p-values indicate statistically significant differences ($p < 0.05$). Extremely small p-values (< 0.0001) are displayed as `0.0000` for readability and represent highly significant effects rather than numerical zero.

4.1 Unimodal Audio

Table 1 reports the validation performance of audio-based classifiers using ViT, CNN, and HuBERT backbones. The ViT-tiny model achieved the highest F1 score and balanced accuracy, while HuBERT obtained the highest recall.

Model	F1	Bal. Acc.	Prec.	Rec.
ViT(-tiny)	**0.7654**	**0.6510**	**0.6327**	0.9688
CNN(-custom)	0.7373	0.6260	0.6214	0.9062
HuBERT	0.7287	0.5698	0.5802	**0.9792**

Table 1: Performance of unimodal audio models.

Pairwise comparisons in Table 2 show that ViT differs significantly from both CNN and HuBERT across most metrics.

Comparison	F1	Bal. Acc.	Prec.	Rec.
ViT vs CNN	**0.0000**	**0.0000**	**0.0000**	**0.0000**
ViT vs HuBERT	**0.0000**	**0.0000**	**0.0000**	0.4353
CNN vs HuBERT	0.0868	**0.0000**	**0.0000**	**0.0000**

Table 2: Pairwise Welch's unequal-variance t-test (p-values) between unimodal audio models.

4.2 Unimodal Text

Table 3 summarizes the validation performance of text-based models using BERT and DistilBERT under ARPA and non-ARPA transcription variants. BERT with ARPA transcription achieved the joint-highest F1 and balanced accuracy (tied with DistilBERT on non-ARPA), while DistilBERT performed comparably across both representations.

Configuration	F1	Bal. Acc.	Prec.	Rec.
BERT (ARPA)	**0.7287**	**0.5922**	**0.5960**	0.9375
BERT (non-ARPA)	0.7111	0.5185	0.5517	**1.0000**
DistilBERT (ARPA)	0.7200	0.5737	0.5844	0.9375
DistilBERT (non-ARPA)	**0.7287**	**0.5922**	**0.5960**	0.9375

Table 3: Performance of unimodal text models.

Pairwise comparisons in Table 4 show significant differences between BERT and DistilBERT under non-ARPA conditions, whereas ARPA-based variants were largely comparable.

Comparison	F1	Bal. Acc.	Prec.	Rec.
BERT: ARPA vs non-ARPA	**0.0000**	**0.0000**	**0.0000**	**0.0013**
DistilBERT: ARPA vs non-ARPA	**0.0000**	**0.0055**	**0.0159**	0.9553
BERT vs DistilBERT (non-ARPA)	**0.0000**	**0.0000**	**0.0000**	**0.0000**
BERT vs DistilBERT (ARPA)	0.6652	0.1206	0.1028	**0.0419**

Table 4: Welch's unequal-variance t-tests (p-values) for ARPA vs. non-ARPA and BERT vs. DistilBERT.

4.3 Multimodal Audio–Text Fusion

Table 5 presents the best multimodal configurations across all backbones. Among them, CNN–BERT (non-ARPA) achieved the highest balanced accuracy, precision, and F1.

Audio Encoder	Text Encoder	ARPA	F1	Bal. Acc.	Prec.	Rec.
ViT	BERT	No	0.7531	0.6416	0.6294	0.9375
ViT	BERT	Yes	0.7364	0.6188	0.6154	0.9167
CNN	BERT	No	**0.7615**	**0.6530**	**0.6364**	**0.9479**
HuBERT	BERT	No	0.7154	0.5756	0.5867	0.9167

Table 5: Best multimodal configurations across all backbones.

Welch's unequal-variance t-test results in Table 6 reveal that ARPA versus non-ARPA differences are significant primarily for ViT- and CNN-based architectures.

Audio Encoder	Text Encoder	F1	Bal. Acc.	Prec.	Rec.
ViT	BERT	**0.0000**	**0.0000**	**0.0000**	0.3895
ViT (Arch)	BERT	**0.0319**	**0.0221**	**0.0494**	0.7238
CNN	BERT	**0.0000**	**0.0179**	0.0728	**0.0000**
HuBERT	BERT	0.1170	0.4832	0.8879	0.5993

Table 6: Welch's unequal-variance t-tests (p-values) for ARPA vs. non-ARPA multimodal models, all using BERT as text encoder.

Table 7 compares BERT and DistilBERT encoders within multimodal setups. Most variants exhibit significant differences, particularly in ViT- and CNN-based configurations.

Model	ARPA	F1	Bal. Acc.	Prec.	Rec.
ViT	Yes	**0.0000**	**0.0000**	**0.0000**	0.1265
ViT	No	0.8343	**0.0000**	**0.0000**	**0.0000**
ViT (Arch)	Yes	**0.0002**	0.6331	0.7432	**0.0064**
ViT (Arch)	No	0.1289	**0.0003**	**0.0000**	**0.0024**
CNN	Yes	0.8827	**0.0001**	**0.0001**	**0.0303**
CNN	No	0.8838	0.9974	0.9765	0.8659
HuBERT	Yes	0.3558	**0.0011**	**0.0050**	**0.0294**
HuBERT	No	0.1371	**0.0000**	**0.0000**	**0.0055**

Table 7: Welch's unequal-variance t-tests (p-values) for BERT vs. DistilBERT in multimodal configurations.

Table 8 contrasts ViT–BERT backbones under ARPA and non-ARPA text conditions. The ViT-Small and ViT-Tiny models show marginal yet significant F1 differences between transcription variants.

Backbone	Text Encoder	F1 (ARPA)	F1 (non-ARPA)	Δ	p
ViT-Base	BERT	0.7116	0.7169	-0.0052	0.2927
ViT-Small	BERT	0.7190	0.7078	$+0.0112$	**0.0410**
ViT-Tiny	BERT	0.7243	0.7276	-0.0033	**0.0469**

Table 8: Comparison across ViT backbones (architecture search), all using BERT as text encoder.

4.4 Comparison Across Modalities

Table 9 summarizes the best-performing configurations across unimodal and multimodal setups. The unimodal ViT-tiny model attained the highest overall F1, while the multimodal CNN–BERT configuration yielded the best balanced accuracy and precision.

Configuration	F1	Bal. Acc.	Prec.	Rec.
Unimodal Audio (ViT)	**0.7654**	0.6510	0.6327	**0.9688**
Unimodal Text (BERT, ARPA)	0.7287	0.5922	0.5960	0.9375
Multimodal (ViT + BERT, non-ARPA)	0.7531	0.6416	0.6294	0.9375
Multimodal (CNN + BERT, non-ARPA)	0.7615	**0.6530**	**0.6364**	0.9479
Multimodal (HuBERT + BERT, non-ARPA)	0.7154	0.5756	0.5867	0.9167

Table 9: Best results across unimodal and multimodal configurations.

Across all settings, the unimodal ViT model confirmed the strong discriminative capacity of spectro–temporal features, whereas the multimodal CNN–BERT (non-ARPA) system provided superior balanced accuracy and precision. A direct Welch's unequal-variance t-test between these two top-performing models indicated a small but significant difference ($p < 0.001$), suggesting complementary rather than strictly superior information between modalities.

Held-Out Test Evaluation. Evaluation on the held-out test set confirmed the generalization performance of the retrained unimodal and multimodal configurations. Test results closely mirrored the validation patterns, with an average F1 decrease of approximately 0.1. Validation–test correlations confirmed this consistency, showing significant associations for balanced accuracy and precision ($r \approx 0.6$–0.65, $p < 0.05$) and non-significant associations for F1 and recall ($r \approx 0.1$–0.2, $p > 0.5$).

5 Discussion

This section interprets the experimental results presented in Section 4, examining model-specific trends, modality interactions, and the clinical relevance of observed performance patterns.

5.1 Unimodal Audio Models

The ViT encoder significantly outperformed CNN and HuBERT on F1 and balanced accuracy (Table 1), supporting the hypothesis that transformer-style patch embeddings capture longer-range spectro–temporal dependencies in audio representations [13, 17].

HuBERT achieved the highest recall but lower precision and balanced accuracy, reflecting a typical sensitivity–specificity trade-off. This pattern may correspond to the slowed but articulatorily intact speech patterns often observed in MCI [2, 12, 49].

CNN outperformed HuBERT in balanced accuracy and precision but achieved lower recall (Table 2), suggesting that convolutional representations provide more specific but slightly less sensitive discrimination of pathological cues.

5.2 Unimodal Text Models

ARPA transcriptions improved F1 and balanced accuracy for BERT, while DistilBERT performed best on non-ARPA text (Table 3). Significance tests revealed differences for BERT across all metrics and for DistilBERT across F1, balanced accuracy, and precision (Table 4). BERT and DistilBERT performed similarly under ARPA conditions but diverged significantly on non-ARPA text. KDE and violin plots indicated that ARPA reduced performance dispersion and stabilized tokenization-sensitive models. Overall, the text branch achieved strong recall but lagged behind audio models in balanced accuracy.

The gain from ARPA likely reflects its explicit phonemic tokenization, which aligns more closely with transformer subword embeddings than orthographic text [25, 47]. DistilBERT's compressed architecture [44] may limit its ability to exploit such fine-grained phonetic cues, explaining its comparatively better performance on orthographic inputs. This contrast suggests that transcription granularity interacts with model capacity, suggesting that phonemic abstraction benefits larger encoders but not distilled variants.

5.3 Multimodal Audio–Text Fusion

Given the limited benefit of phonetic ARPAbet transcriptions for DistilBERT in multimodal configurations (see Table 7), DistilBERT was excluded from further multimodal comparisons.

Cross-attention fusion with CNN–BERT (non-ARPA) achieved the highest balanced accuracy and precision overall, with F1 comparable to the best unimodal ViT(-tiny) (Table 5). ARPA versus non-ARPA effects were significant for ViT- and CNN-based BERT fusion (Table 6) but largely non-significant for HuBERT, indicating that transcription choice interacts with the audio backbone. These results suggest that the contribution of phonetic information depends on architectural characteristics: measurable for spectrogram-based models but negligible for waveform-level encoders. Within ViT–BERT variants, small yet significant F1 differences between ARPA and non-ARPA were observed for the small and tiny configurations but not for base, suggesting mild sensitivity to model capacity and patch scale (Table 8).

The superior performance of CNN–BERT compared to ViT–BERT may result from CNNs' smoother, lower-dimensional feature maps, which align more effectively with text embeddings during cross-attention. In contrast, ViT's patch-based representations, though rich in local detail, may introduce redundancy that complicates alignment at small sample sizes [13, 50].

Overall, these findings may indicate that multimodal fusion leverages complementary acoustic and linguistic cues, capturing lexical diversity, narrative cohesion, and syntactic simplification [10, 32, 48]. Together, these cues may help stabilize decision boundaries and improve discrimination in borderline cases. The effectiveness of phonetic transcriptions, which bridge audio and text by encoding sub-lexical units aligned with spectro–temporal structure, appears to depend on the representational overlap and abstraction level of the involved encoders [6, 21, 25, 47].

Comparing Top Systems. A direct Welch's unequal-variance t-test between the best unimodal and best multimodal configuration confirmed their complementarity rather than dominance ($p < 0.001$), indicating that multimodal fusion contributed additional but not superior diagnostic information.

Held-Out Test Set Generalization. Evaluation on the held-out test set was consistent with the validation-based trends: the best-performing unimodal and multimodal configurations preserved their relative ranking, with only a modest F1 decrease consistent with expected generalization gaps on small clinical datasets. Pairwise Welch's unequal-variance t-tests confirmed that the observed performance differences across model fami-

lies and transcription variants remained statistically significant on the held-out test data ($p < 0.05$), indicating that validation metrics provided reliable and robust estimates of true model performance and that the framework generalizes consistently beyond the validation split.

Correlation analyses further revealed strong associations for balanced accuracy and precision and weaker, non-significant relationships for F1 and recall. This consistency supports the robustness of the observed trends, confirming the superior performance of ViT and CNN backbones and the stabilizing effect of ARPA-based transcriptions for BERT.

5.4 Explainability and Clinical Interpretability

To enhance clinical interpretability, segment-level explanations were generated using LIME [41] on log-Mel spectrogram inputs. All explanations were derived from single-channel log-scaled Mel-spectrograms (128 Mel filters, 16 kHz sampling rate) corresponding to unaugmented validation samples. During the explanation process, low-energy and padding regions were masked to reduce artifacts introduced by LIME perturbations, and color overlays were added solely for visualization without affecting inference. Despite this masking, some highlighted regions occasionally appear near padding areas, indicating that the explanation method or model may attribute importance to silent or low-energy regions.

Figure 3 illustrates the corresponding LIME explanations for an HC participant, with the same utterance visualized under HC and MCI prediction perspectives. The explanations appeared to reveal distinct spectro–temporal patterns for HC and MCI predictions. For HC, the model appeared to rely on stable harmonic stacks and regular voiced–unvoiced transitions, corresponding to smooth articulation and rhythmic prosody. For MCI, explanations tended to emphasize spectrally diffuse and temporally unstable regions with blurred harmonic boundaries, consistent with reduced articulatory precision and weakened prosodic regularity [10, 32, 49]. Although classifier confidence was generally higher for HC samples, the contrast between LIME maps revealed interpretable and clinically meaningful distinctions, suggesting that the model captured relevant acoustic cues rather than spurious correlations.

Text-based LIME analyses produced inconsistent importance maps due to tokenization variability and limited sample size; therefore, only acoustic explanations were retained for qualitative interpretation.

5.5 Limitations and Future Work

Data Limitations and Architectural Constraints. The dataset remains relatively small (64 MCI and 48 HC participants), limiting statistical power and generalizability. This constraint is amplified for multimodal architectures, which typically require substantially larger datasets to learn stable cross-modal representations [6]. Given this limitation, the framework was deliberately designed to remain compact and interpretable (e.g., single-head cross-attention, lightweight classifiers) to mitigate overfitting, which may reduce representational capacity compared to larger multimodal transformers.

Statistical Reliability and Evaluation. All main results were obtained using a single random seed for systematic reporting, with exploratory reruns for verification. While Welch's unequal-variance *t*-tests provide informative comparisons under fixed participant splits, they do not replace uncertainty estimates at the participant level. Future work should therefore employ repeated experiments with multiple seeds or *k*-fold cross-validation [23] and report confidence intervals (e.g., via bootstrap over participants [14]) to obtain more stable estimates.

Explainability Limitations. LIME approximates model behavior through local perturbation-based surrogates [41], which may not fully capture global decision boundaries and can yield unstable explanations in high-dimensional spaces [3]. Text-based LIME produced highly variable patterns, preventing reliable interpretation; thus only acoustic explanations were reported. Interpretability results should therefore be viewed as exploratory and complementary rather than definitive evidence of model reasoning.

Linguistic Representation. The ARPAbet transcription scheme encodes coarse phonemic categories and lacks the fine-grained articulatory detail of the International Phonetic Alphabet (IPA) [5]. Extending preprocessing to IPA or hybrid grapheme–phoneme representations could improve phonetic resolution and enhance interpretability.

Future Directions. Beyond acoustic–lexical fusion, future research should incorporate discourse-level features of the CTP task, such as spatial organization, sequencing, and referential cohesion-markers of executive control and cognitive planning

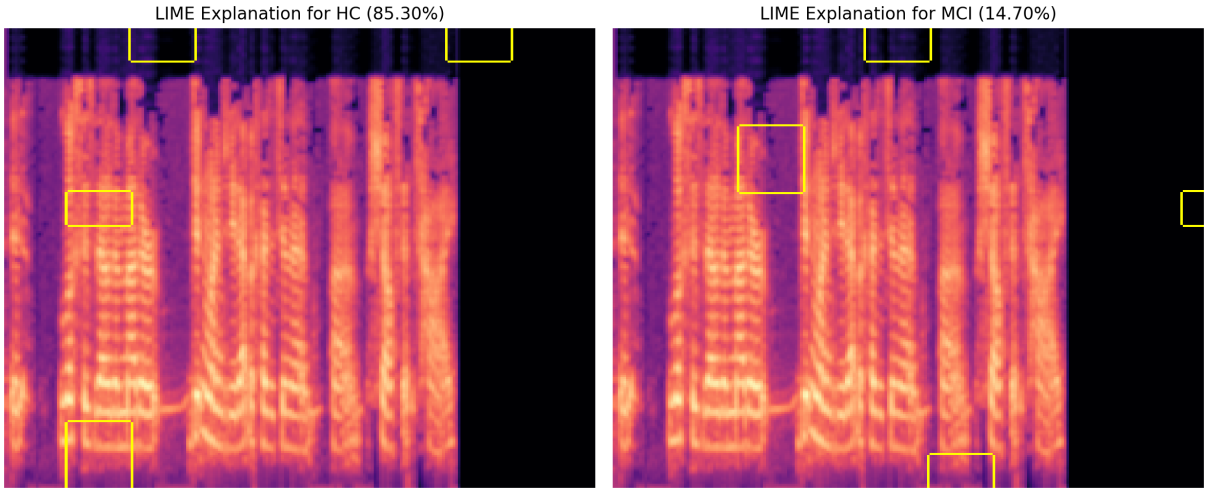

Figure 3: LIME explanations for a ViT-based log-Mel spectrogram from an HC participant, highlighting regions that show evidence for HC (left) and MCI (right). Yellow boxes denote the most influential time–frequency patches.

in MCI [15]. Extending the framework to spontaneous or conversational speech (e.g., human–machine interactions) would further improve ecological validity. Finally, real-world deployment will require standardized protocols for data acquisition and task administration to ensure reproducibility and clinical applicability.

6 Conclusion

This study presented a unified multimodal framework for classifying MCI versus HC using synchronized speech and text from the Delaware corpus within DementiaBank. The framework integrates audio and text encoders via cross-attention fusion and supports flexible unimodal and multimodal configurations under consistent training, optimization, and evaluation protocols.

Across all modalities, the Vision Transformer (ViT) provided the strongest unimodal audio baseline, confirming the discriminative potential of spectro–temporal representations, whereas BERT-based text models achieved the best balance between precision and recall. Multimodal fusion with CNN–BERT (non-ARPA) yielded slightly higher balanced accuracy and precision than unimodal models, indicating complementary acoustic–linguistic cues that enhance diagnostic consistency across classes. Although multimodal gains were modest, they were statistically significant ($p < 0.001$) and demonstrated that combining speech and language information provides complementary diagnostic value.

Phonetic ARPAbet transcriptions improved the stability and performance of BERT models, particularly within multimodal fusion frameworks, while providing limited benefit for DistilBERT. This indicates that transcription granularity interacts with both model capacity and the nature of audio–text fusion architectures.

Segment-level LIME analyses revealed interpretable spectro–temporal regions associated with articulatory and prosodic irregularities characteristic of early cognitive decline, underscoring the clinical relevance of learned acoustic features.

Overall, the findings highlight the potential of transparent and reproducible multimodal modeling for speech-based cognitive screening. Future work could extend this framework to larger and more diverse datasets, enriched linguistic representations (e.g., IPA or discourse-level features), and standardized evaluation pipelines to enhance reliability and clinical applicability.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, New York, NY, USA. ACM.

[2] Muath Alsuhaibani, Ali Pourramezan Fard, Jian Sun, Farida Far Poor, Peter S. Pressman, and Mohammad H. Mahoor. 2025. A review of machine learning approaches for non-invasive cognitive impairment detection. *IEEE Access*, 13:56355–56384.

[3] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *CoRR*, abs/1806.07538.

[4] Pranav S. Ambadi, Kristin Basche, Rebecca L. Koscik, Visar Berisha, Julie M. Liss, and Kimberly D. Mueller. 2021. Spatio-semantic graphs from picture description: Applications to detection of cognitive impairment. *Frontiers in Neurology*, Volume 12 - 2021.

[5] International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

[6] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP.

[7] Lukas Biewald. 2020. Experiment tracking with weights and biases. `https://www.wandb.com/`. Software available from wandb.com.

[8] Howard Chertkow. 2002. Mild cognitive impairment. *Current Opinion in Neurology*, 15(4):401–407.

[9] Louise Cummings. 2019. Describing the cookie theft picture. *Pragmatics and Society*, 10(2):153–176.

[10] Celine De looze, Amir Dehsarvi, Lisa Crosby, Aisling Vourdanou, Robert Coen, Brian Lawlor, and Richard Reilly. 2021. Cognitive and structural correlates of conversational speech timing in mild cognitive impairment and mild-to-moderate alzheimer's disease: Relevance for early detection approaches. *Frontiers in Aging Neuroscience*, 13:637404.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[12] Valeria Diaz and Guillermo Rodríguez. 2022. Machine learning for detection of cognitive impairment. *Acta Polytechnica Hungarica*, 19:195–213.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

[14] Stephanie Fook Chong and Robin Choo. 2011. Introduction to bootstrap. *Proceedings of Singapore Healthcare*, 20:236–240.

[15] Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422. PMID: 26484921.

[16] Serge Gauthier, Barry Reisberg, Michael Zaudig, Ronald C. Petersen, Karen Ritchie, Karl Broich, Sylvie Belleville, Henry Brodaty, David Bennett, Howard Chertkow, Jeffrey L. Cummings, Mony de Leon, Howard Feldman, Mary Ganguli, Harald Hampel, Philip Scheltens, Mary C. Tierney, Peter Whitehouse, and Bengt Winblad. 2006. Mild cognitive impairment. *The Lancet*, 367(9518):1262–1270.

[17] Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: audio spectrogram transformer. *CoRR*, abs/2104.01778.

[18] F. J. Harris. 1978. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[20] Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

[21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

[22] David S. Knopman and Ronald C. Petersen. 2014. Mild cognitive impairment and mild dementia: A clinical perspective. *Mayo Clinic Proceedings*, 89(10):1452–1459.

[23] Ron Kohavi and 1 others. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

[24] Alyssa M. Lanzi, Anna K. Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L. Cohen. 2023. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

[25] Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023. Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions. *Preprint*, arXiv:2301.08810.

[26] Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

[27] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *Preprint*, arXiv:2004.06833.

[28] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *Preprint*, arXiv:2104.09356.

[29] Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

[30] J. C. Morris, M. Storandt, J. P. Miller, D. W. McKeel, J. L. Price, E. H. Rubin, and L. Berg. 2001. Mild cognitive impairment represents early-stage alzheimer disease. *Archives of Neurology*, 58(3):397–405.

[31] John C. Morris. 2005. Mild cognitive impairment and preclinical alzheimer's disease. *Geriatrics*, Suppl:9–14.

[32] Kimberly D. Mueller, Bruce Hermann, Jonilda Mecollari, and Lyn S. Turkstra. 2018. Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40(9):917–939. PMID: 29669461.

[33] Marko Niemelä, Mikaela von Bonsdorff, Sami Äyrämö, and Tommi Kärkkäinen. 2025. Dementia classification using acoustic speech and feature selection. *Preprint*, arXiv:2502.03484.

[34] RM Ochshorn and Max Hawkins. 2017. Gentle forced aligner. *github. com/lowerquality/gentle*.

[35] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech$_2$019.$ISCA$.

[36] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

[39] Ali Pourramezan Fard, Mohammad H. Mahoor, Muath Alsuhaibani, and Hiroko H. Dodge. 2024. Linguistic-based mild cognitive impairment detection using informative loss. *Computers in Biology and Medicine*, 176:108606.

[40] Lutz Prechelt. 1998. *Early Stopping - But When?*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg.

[41] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

[42] Marwan N. Sabbagh, M. Boada, S. Borson, M. Chilukuri, P. M. Doraiswamy, B. Dubois, J. Ingram, A. Iwata, A. P. Porsteinsson, K. L. Possin, G. D. Rabinovici, B. Vellas, S. Chao, A. Vergallo, and H. Hampel. 2020. Rationale for early diagnosis of mild cognitive impairment (mci) supported by emerging digital technologies. *The Journal of Prevention of Alzheimer's Disease*, 7(3):158–164.

[43] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.

[44] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[45] Arezo Shakeri, Mina Farmanbar, and Krisztian Balog. 2025. Multiconad: A unified multilingual conversational dataset for early alzheimer's detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3627–3637, New York, NY, USA. Association for Computing Machinery.

[46] Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. 2025. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3(2):100205.

[47] Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript. *Preprint*, arXiv:2102.00804.

[48] Vanessa Taler and Natalie A. Phillips. 2008. Language performance in alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556. PMID: 18569251.

[49] Laszlo Toth, Ildiko Hoffmann, Gabor Gosztolya, Veronika Vincze, Greta Szatloczki, Zoltan Banreti, Magdolna Pakaski, and Janos Kalman. 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138.

[50] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *CoRR*, abs/1906.00295.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

[52] B. L. Welch. 1947. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

[53] Ross Wightman. 2019. Pytorch image models.

[54] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[55] Jiawei Zhang. 2019. Cognitive functions of the brain: Perception, attention and memory. *Preprint*, arXiv:1907.02863.

Appendix

A Declaration of Academic Integrity

I confirm that this guided research is my own work and I have documented all sources and material used.

B Declaration on the Use of AI Tools

In preparing and revising this paper/report, I used artificial intelligence tools (such as ChatGPT) for language refinement, structural editing, formatting suggestions, and limited coding assistance. All conceptual, analytical, and experimental work, including data processing, modeling, and interpretation, is my own original contribution. The AI-generated text and code were critically reviewed, verified, and where necessary substantially revised by me to ensure factual correctness and academic integrity. No AI system was used to generate original research content, results, or analyses.