Master's Thesis

Cognitive States Prediction based on Multimodal Analysis in Online Meeting with Coimagination Method

WEI Jingtong

Supervisor Shogo Okada

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2026

Abstract

Mild Cognitive Impairment (MCI) represents a transitional stage between normal aging and dementia and is a key target for early intervention. Conventional clinical assessment typically relies on standardized scales and face-to-face interviews, which can be costly, infrequent, and limited in accessibility. With the rapid growth of online communication, developing non-invasive and scalable screening approaches based on natural conversational behavior has become increasingly meaningful in practice. However, in natural interactive settings, linguistic, acoustic, and facial behaviors are shaped not only by cognitive status but also by mood fluctuations, social strategies, and individual differences. These confounding factors can weaken feature discriminability and undermine model generalization. To address this challenge, this thesis investigates session-level multimodal MCI recognition in an online Coimagination meeting environment, with a particular focus on reducing the confounding effect of mood changes when incorporating facial Action Units (AUs), thereby improving robustness.

In this work, we formulate MCI screening at the session level and construct multimodal representations from language, speech, and video signals, together with facial AU features extracted by OpenFace. For AUs, utterance-level activation is determined using the proportion of activated frames within each utterance time window, and session-level AU descriptors are obtained by aggregating utterance-level binary activations across the session. Audio, text, and video modalities are represented using embeddings from pretrained models, where the video feature dimension is 1024. These modality-specific features are used to form session-level feature vectors for classification. To mitigate mood-related confounding in AU-based modeling, we use the change in Face Scale ratings before and after each session (ΔFace Scale) to quantify mood variation. Rather than coarse grouping, each distinct ΔFace Scale value is treated as a stratum. Within each stratum, we perform significance testing between the MCI group and the healthy control group on AU features, and select statistically significant AUs to form a reduced AU subset for subsequent classification experiments.

For classification, we implement a linear-kernel SVM as a conventional machine-learning baseline and a multi-branch DNN as the deep model. In the DNN, each modality (AU, audio, text, and video) is encoded by an independent MLP branch to learn compact latent representations, followed by feature-level fusion at an intermediate layer through concatenation and a final classification head. To ensure strict evaluation and fair comparability,

both SVM and DNN adopt Leave-One-Group-Out (LOGO) cross-validation, where each participant (all sessions of a subject) defines a group, enabling evaluation of generalization to unseen participants. Performance is reported using Accuracy and Macro F1, with systematic ablation over unimodal and multimodal settings. Since AU and video are both facial-information channels and may introduce redundancy or instability in fusion learning, AU and video are not combined within the same experimental setting.

Experimental results indicate that multimodal fusion benefits the deep model more substantially, and that mood-informed AU screening affects SVM and DNN differently. For configurations without AU, the best DNN performance is achieved by Audio+Text fusion (Acc = 0.725, Macro F1 = 0.683), outperforming the corresponding linear SVM and highlighting the advantage of deep feature fusion for capturing complementary speech–language cues. Using the full (unscreened) AU set, the DNN achieves the overall best performance under AU-only (Acc = 0.742, Macro F1 = 0.675), suggesting that the deep model can exploit nonlinear discriminative patterns from richer AU representations; AU-only is also the strongest setting for SVM under this condition, though with lower overall performance than DNN. After applying ΔFace Scale–guided screening, SVM improves under AU-only (Acc = 0.710, Macro F1 = 0.642), indicating that stratified significance-based selection effectively reduces mood confounds and weakly discriminative AUs for a linear classifier. In contrast, the best DNN setting with screened AUs shifts to AU+Text (Acc = 0.711, Macro F1 = 0.681), implying that a more compact AU subset can be more effectively complemented by linguistic information in deep multimodal learning.

In summary, this thesis presents a reproducible session-level multimodal framework for MCI recognition in interaction scenarios of online meeting scenarios and demonstrates that ΔFace Scale–guided stratified AU screening can mitigate mood-related confounding and improve robustness. The proposed methodology and empirical findings provide practical evidence for non-invasive cognitive screening in natural interaction settings and offer a foundation for future studies with larger-scale and multi-site online screening data.

Keywords: Mild Cognitive Impairment (MCI), Multimodal Learning, Online Conversation Analysis, Cognitive Screening

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mild Cognitive Impairment(MCI) is a degenerative disorder primarily characterized by the progressive deterioration of cognitive functions, including memory, communication, and judgment. As the condition advances, individuals often experience significant difficulties in daily activities, social interactions, and informed decision-making. According to recent global estimates, approximately 50 million people currently live with MCI, and a new case is diagnosed every three seconds. With continued population aging and improved diagnostic practices, the total number of individuals with MCI is projected to triple by 2050[1]. Despite extensive research endeavors, there remains no definitive cure for the disease. However, by detecting the early stages of MCI, timely interventions can be implemented to mitigate the associated symptoms, thereby improving individuals' quality of life. Given the rising prevalence of this disorder, developing novel diagnostic tools and therapeutic strategies to slow its progression and offer more robust support to patients and their caregivers has become increasingly urgent. Notably, the rapid advances in artificial intelligence (AI) technologies, including multimodal data analysis, machine learning, and deep learning, hold significant promise for facilitating earlier detection and intervention[2].

Within this evolving landscape of MCI research, the Coimagination Method has emerged as a particularly effective technique for identifying mild cognitive impairment (MCI)[3][4]. It is a conversation-based method involving a group of participants, typically four and one host. The host engages in structured discussions centered on prepared images or themes. Each session is divided into two phases. In the first phase, each participant presents one or two prepared images within a strict one-minute timeframe, fostering organized thought and encouraging focused attention. In the second phase, the other participants pose questions, and the speaker responds within approximately two minutes per query. Throughout this process, the host observes partici-

pant interactions and ensures balanced engagement by prompting individuals who do not pose questions.

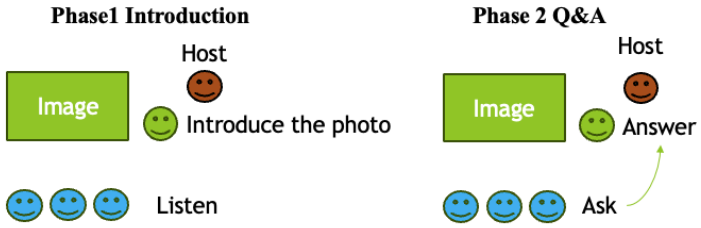

Figure 1.1: Coimagination Method

This structured format promotes active listening dialogue, enabling researchers to detect subtle signs of cognitive decline, such as alterations in comprehension, recall, and verbal fluency. The Coimagination method has been successfully implemented in community settings, serving as a valuable, group-based approach for the early detection of MCI. By fostering a supportive environment encourages meaningful interaction and helps uncover cognitive deficits, thereby facilitating timely intervention and potentially improving clinical outcomes[3][4].

With the growing popularity of online meetings, the Coimagination method now has a novel medium through which it can be implemented. Traditionally, conducting Coimagination method sessions requires assembling participants in person, which is both time-consuming and resource-intensive due to the need for a physical site and specialized equipment. By contrast, an online meeting format offers a contactless approach that can reduce both the duration and cost of each assessment, thereby shortening the detection cycle. Moreover, online conference platforms generate multimodal data—such as video, audio, and textual content—that can be leveraged by artificial intelligence models to facilitate more efficient and accurate feature extraction. Consequently, applying the Coimagination Method in an online setting holds promise for enhancing the accessibility, scalability, and cost-effectiveness of early MCI detection.

Beyond linguistic and acoustic cues, facial behaviors provide an additional window into participants' internal states during interaction. Prior studies suggest that both mood (affective state) and cognitive state may manifest through facial action units (AUs), and that the cognitive changes associated with MCI can be reflected in AU activation patterns. However, AU expressions are not determined by cognitive status alone. Contextual factors such as interpersonal or human-computer interaction dynamics, the participant's

current affective baseline, and subjective feelings during the conversation can also influence mood and, consequently, the observed AUs[5]. In our dataset, mood is operationalized by Face Scale ratings recorded before and after each session, which serve as a proxy of affective state. Therefore, to more reliably investigate the relationship between AU patterns and MCI-related cognitive impairment, this study treats mood (Face Scale or its change) as a potential confounding factor and aims to control its influence, so that AU-based evidence of cognitive decline can be examined with reduced affect-driven variability.

Building on these insights, this study proposes a computational framework intended to identify individuals who may be in the early stages of MCI by examining their interactions in online Coimagination sessions. Through this approach, the research aims to enrich clinical practice by offering a contactless, large-scale MCI screening method that alleviates the constraints of in-person testing. By integrating multimodal deep learning techniques and emphasizing linguistic, acoustic, and visual signals, the proposed system seeks to uncover cognitive deficits that might otherwise remain unrecognized. Moreover, by extending the Coimagination method to online meeting platforms, this work not only addresses a critical gap in current MCI detection approaches but also provides a cost-effective and time-efficient solution. By this method, it has the potential to streamline diagnostic workflows, reduce healthcare burdens, and offer timely interventions for individuals exhibiting early signs of cognitive decline.

Chapter 2

Related Works

This study focuses on predicting cognitive states in online meetings with the Coimagination Method[3] based on multimodal analysis. At present, there is a lack of research on MCI by analyzing people's social status, but there are studies on MCI by analyzing related multimodal patterns. We hypothesize that machine learning techniques can be applied to the online meeting format of the Coimagination method to identify states indicative of MCI risk.

2.1 MCI analysis method

MCI analysis methods can be broadly categorized into neuropsychological assessments, neuroimaging techniques, and AI-driven multimodal analysis. Traditional neuropsychological assessments, such as the Mini-Mental State Examination (MMSE)[6] and Montreal Cognitive Assessment (MoCA)[7], evaluate cognitive functions through structured cognitive tasks. While these methods are widely used for clinical diagnosis, their accuracy is influenced by education level, cultural background, and examiner variability, making them less reliable for early-stage MCI detection.

Neuroimaging techniques, including MRI, PET, and fMRI, provide insights into structural and functional brain changes associated with MCI[8][9][10]. MRI is commonly used to detect hippocampal atrophy, while PET scans reveal amyloid-beta accumulation, a hallmark of Alzheimer's disease[9]. However, these methods require expensive equipment and specialized expertise, limiting their accessibility for large-scale screening.

Recent advancements in AI-driven multimodal MCI analysis integrate data from speech, text, facial expressions, and physiological signals to enhance diagnostic accuracy[2]. Studies have demonstrated that natural language processing (NLP) models can analyze linguistic patterns in sponta-

neous speech to detect cognitive decline. Similarly, computer vision techniques can assess facial expressions and eye movements to identify behavioral markers of MCI. Multimodal approaches that combine these features with machine learning algorithms have shown promising results in early-stage detection.

Given the limitations of traditional methods, the integration of multimodal AI-based analysis offers a more scalable, objective, and cost-effective approach to MCI detection.

2.1.1 Analysis of speech and language patterns

Speech and language are among the most mature and evidence-rich behavioral cues for the early identification of MCI. Prior studies indicate that impairments in attention, memory, and executive function lead to language decline[11], motivating a growing body of work that combines NLP with machine learning[12]. Common elicitation paradigms include constrained tasks (reading, repetition)[13][14], unconstrained tasks (spontaneous conversation)[15], verbal fluency tasks (e.g., animal naming)[13][16], picture description (e.g., Cookie Theft Picture Description Task)[17][13][16], and cognitively demanding tasks such as serial subtraction. In particular, free conversation often separates patients more effectively than constrained tasks because it relies more on sentence planning and spontaneous lexical retrieval[16], revealing word-finding difficulties earlier. At the feature level, MCI patients frequently show reduced information density, imprecise word choice, increased use of pronouns/articles, lexical disfluencies such as repetition, disrupted syntax and discourse organization, and changes in utterance length; meanwhile, acoustic/prosodic and temporal measures (pause frequency/duration, speech rate/clarity, proportion of silence, etc.) have been repeatedly validated as predictive and can yield strong discrimination in interview or conversational data[18][19][20][21][22].

2.1.2 Analysis of eye gaze

Eye gaze is regarded as a highly sensitive "window" into early impairment[23] because it closely reflects mechanisms such as attention, inhibitory control, and executive function. Subtle eye-movement abnormalities are difficult to observe in routine clinical practice; however, advances in infrared eye trackers and computer-vision-based iris localization enable non-invasive quantification of gaze instabilit[24], increased blinking, impaired eye–head/eye–hand coordination, abnormalities in reflexive and voluntary saccades, and longer

latency and higher error rates in anti-saccade tasks. These measures are typically derived from saccade/anti-saccade paradigms[25], reading tasks, image/video viewing, or visual recognition memory tasks, and have been used to detect early decline in MCI and mild stages[23][24][26]. Overall, eye-gaze research is less mature than speech or gait, but its rich measurement space and early sensitivity make it promising; more robust quantitative analysis is still needed to improve differentiation from related conditions such as Parkinson's disease.

2.1.3 Analysis of facial expressions

Facial-expression studies suggest that although expressive ability may be relatively preserved in early stages[27], measurable differences in affect and facial movements can already emerge, such as a stronger tendency toward negative expressions, changes in lip movements and blinking, and increased eye-closing or mouth-opening under discomfort[28][29][30]. Research covers both spontaneous expressions and responses elicited in clinical pain contexts or during interactions with virtual agents. Some studies leverage FACS[31] and tools such as OpenFace[32] to extract features including mouth-related movements, gaze angles, head pose, and action-unit intensities for diagnostic modeling. Nevertheless, this direction faces notable bottlenecks[33][34]: landmark detection can be biased in clinical dementia populations (for both frontal and profile views), and the lack of large public datasets limits the generalization of facial-expression and emotion-recognition models—making the area promising but still under-explored.

2.1.4 Combined multimodal analytics

Multimodal fusion has become a major trend in the identification of MCI. Prior work most commonly combines mature modalities such as speech, language, and gait, and some studies additionally incorporate eye gaze and facial expression[18][35]; in neuroimaging, EEG, MRI, and PET are often combined to improve diagnostic accuracy. Comparative analyses in the review indicate that multimodal approaches achieve an average improvement of about 15.17% over the "best single modality," highlighting the consistent value of cross-modal complementarity for robustness. To enhance generalization and noise tolerance, key design principles include increasing the number of fused modalities and ensuring heterogeneity and complementarity, so that more evidential sources are available across varying real-world conditions—providing a clear engineering path toward practical MCI screening systems.

2.2 Coimagination method

MCI cognitive support systems have been widely studied, with various approaches focusing on either physiological or cognitive interventions. The Coimagination Method proposed by Otake et al. (2009)[3] is a communication-based cognitive training method aimed at delaying cognitive decline by facilitating structured discussions using collected images. Unlike traditional reminiscence therapy, which primarily focuses on recalling past experiences, this method encourages participants to share emotions and engage in discussions about the past, present, and future. The approach targets the activation of episodic memory, divided attention, and planning ability, which are known to decline in individuals with mild cognitive impairment (MCI).

The study conducted by Otake et al. implemented weekly Coimagination sessions over five weeks in a welfare institution, where elderly participants engaged in structured discussions based on given themes. A memory task was conducted in the final session to assess retention and cognitive engagement. The results demonstrated significant improvements in memory recall and social interaction, supporting the method's effectiveness in cognitive stimulation.

Furthermore, this research identified several cognitive markers relevant to early MCI detection. Participants exhibiting memory recall deficiencies struggled to remember image owners and discussion themes, suggesting episodic memory impairment. Attention deficits were observed in individuals who had difficulty following discussions or recalling multiple interactions. Planning difficulties were evident in those who struggled to select and organize images coherently. Additionally, social disengagement was noted in participants with minimal interaction and responsiveness. These characteristics indicate that the Coimagination Method can serve not only as an intervention for cognitive decline but also as a potential assessment tool for identifying early signs of MCI.

This study provides valuable insights into non-pharmacological MCI interventions and aligns with recent research on multimodal approaches to cognitive health. Given the growing interest in multimodal detection and personalized intervention systems, integrating structured communication-based cognitive stimulation with machine learning-based assessment could further enhance early detection.

2.3 MCI estimation based on Coimagination method

Estimating mild cognitive impairment (MCI) using multimodal analysis has gained increasing attention, particularly in natural conversational settings. Li et al. (2024)[4] proposed an automatic MCI estimation method based on group conversations in the Coimagination Method, leveraging linguistic and acoustic features to classify participants as either MCI or healthy. Unlike traditional MCI detection approaches that focus on individual speech analysis, this study examines multimodal behaviors in interactive group conversations, providing richer contextual cues for cognitive assessment.

The study uses a Japanese Coimagination dataset collected in 2018, containing nine groups of four participants each, where MCI labels were assigned based on MMSE and MoCA scores. Two conversational phases were analyzed separately: (1) the introduction phase, where participants describe their photos, and (2) question-answering phase, where participants actively engage in discussions. Findings indicate that the question-answering phase is more effective for MCI estimation, as it requires greater cognitive engagement.

A deep neural network (DNN) model was employed using linguistic (BERT embeddings, part-of-speech, LIWC) and acoustic (eGeMAPS, Wav2Vec2.0) features. The best multimodal model achieved a macro F1-score of 0.693, demonstrating that turn-taking behavior and acoustic signals are effective in detecting cognitive impairment. Moreover, multitask learning with conversation customary as a subtask improved estimation accuracy, highlighting the importance of behavioral patterns in MCI classification.

This study contributes to showing the effectiveness of multimodal conversational analysis. Future work aims to refine model architectures and explore additional behavioral features to enhance early detection of MCI.

Chapter 3

Dataset

3.1 Dataset of Coimagination Method

In this study, we use a Japanese Coimagination method online meeting dataset collected by RIKEN. The data set includes 31 participants, each of whom participated in five experimental sessions. Each session involved four participants, resulting in a total of 40 videos of online meetings. Each video is approximately 60 minutes in length, recorded at a resolution of 1280×720 and a frame rate of 25 frames per second (fps). Among the 31 participants, showed a tendency toward MCI, as their Mini-Mental State Examination (MMSE) scores were less than 27. We treat participants in each session as independent participants to increase the number of data samples. In addition, this dataset includes the conversation content of the participants in each video and the corresponding time stamp.

3.2 Labels and distribution

MCI labels are annotated based on the threshold of MMSE. The selection of MMSE threshold scores across different studies. We use the threshold of 27 as it is commonly defined the clinical spectrum of MCI. Accordingly, MMSE scores$\leq$ 26 are considered as MCI, and 27-30 is considered as healthy. In this study, if one participant's MMSE score is under the threshold for at least one time, we annotate the participant as MCI, otherwise healthy.

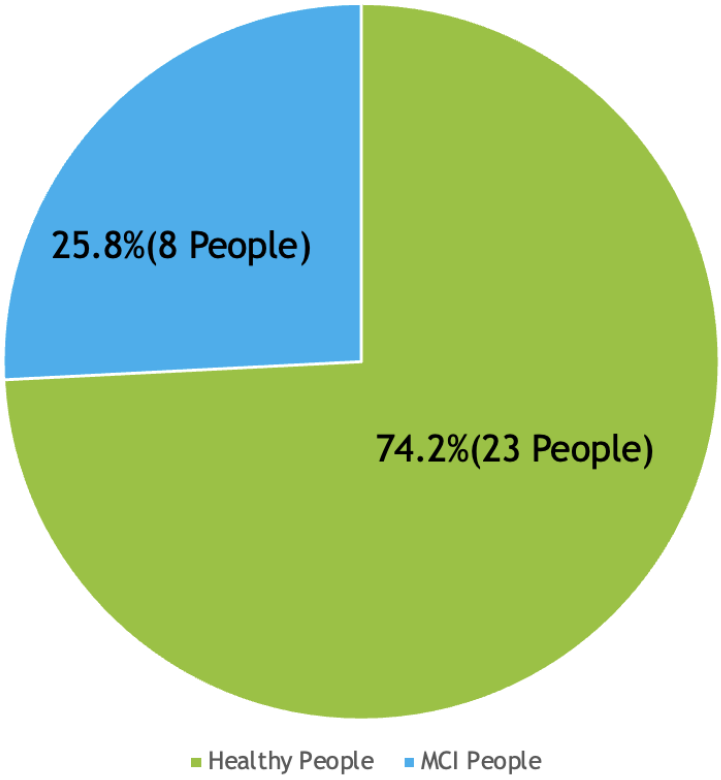

Figure 3.1: Distribution of MCI and Healthy People

We segmented the video and audio streams according to the start and end timestamps of each utterance and ensured temporal consistency across modalities. After segmentation, each modality contained 13,582 items. The relationship between these items and the MMSE scores is shown in the figure below.

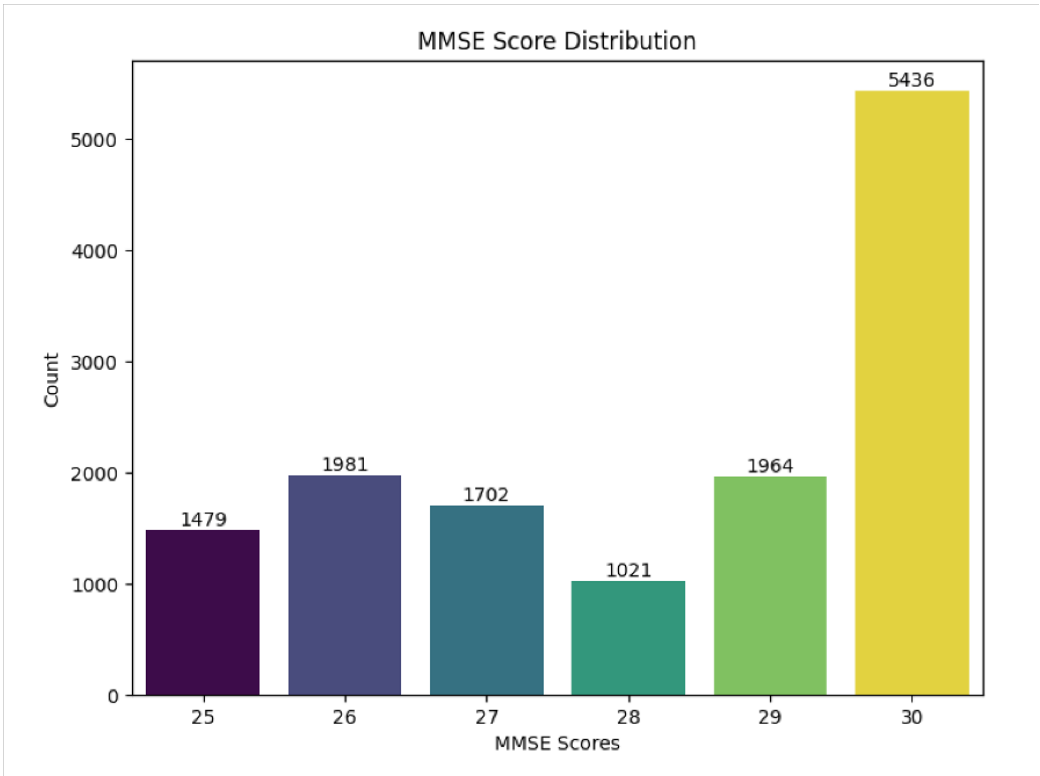

Figure 3.2: MMSE Score Distribution

This dataset uses the **Face Scale** as the **mood label** for each participant. For every experimental session, each participant's mood score is recorded both before and after the session. The mood score consists of seven levels ranging from 6 to 0, spanning from "Best Mood State" to "Worst Mood State".

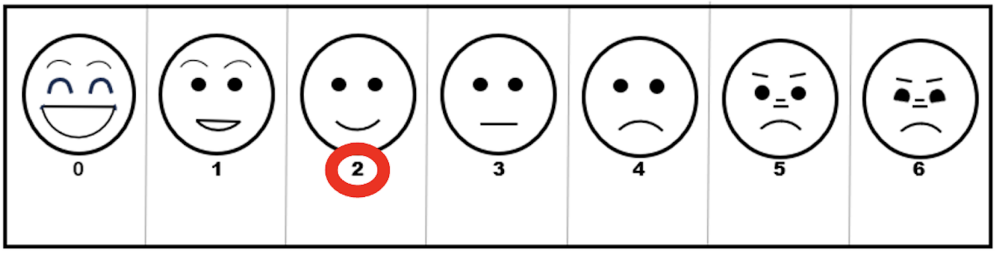

Figure 3.3: Face Scale

To quantify session-level affective variation, mood change was defined using Face Scale ratings as ΔFace_Scale = $Face\_Scale_{after} - Face\_Scale_{before}$

and sessions were categorized into mood improvement (ΔFace_Scale<0) or mood deterioration (ΔFace_Scale>0).

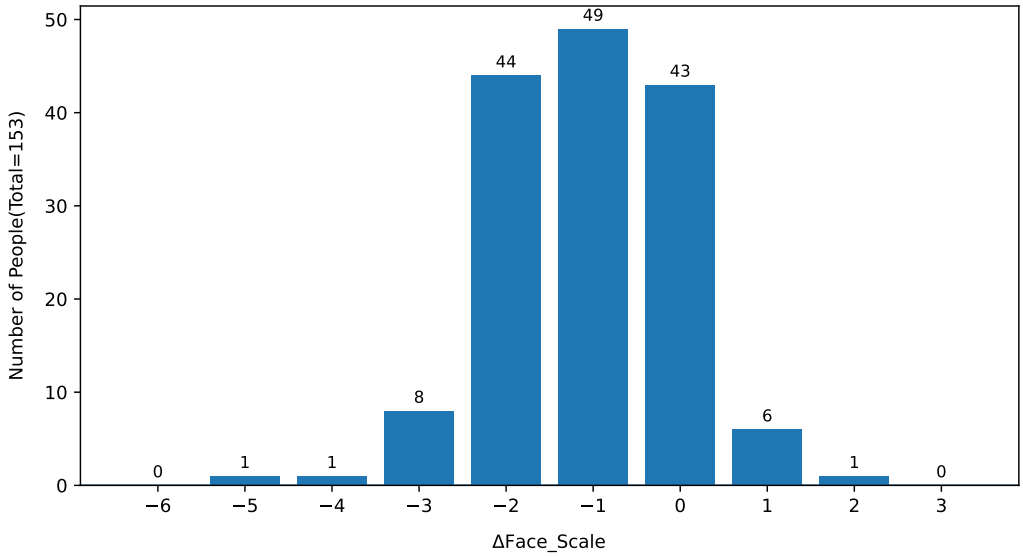

Figure 3.4: Distribution of ΔFace_Scale

Chapter 4

Method

4.1 Framework overview

In this chapter, we propose a multimodal machine learning framework designed to detect individuals' cognitive states by analyzing online Coimagination method meeting environments. By systematically integrating diverse data sources, including speech content, visual cues, and acoustic features, this approach aims to identify potential signs of mild cognitive decline.

A critical methodological challenge addressed in this framework is the inherent ambiguity of non-verbal behavioral cues. Specifically, facial expressions often present a "many-to-one" mapping problem: a single visual indicator, such as a frown, could indicate cognitive difficulty (a symptom of MCI) or simply a transient negative mood. To address this, we introduce a "Mood-Controlled Feature Selection" pipeline to disentangle cognitive markers from emotional variability.

As illustrated in the system architecture (Figure 4.1), the proposed method processes the data through four concrete stages:

1. **Data Processing:** Precise segmentation of audiovisual streams based on speaker timestamps.

2. **Feature Extraction:** Extraction of high-dimensional embeddings utilizing specific pre-trained models, namely Dino v2-large (visual), mms-300m (audio), and Llama 3.1-8b (text), and the facial muscle action units (AUs) were extracted using OpenFace 2.2.

3. **Mood-controlled AUs Selection:** We stratified participants into different mood-change states using the difference between their pre- and post-experiment mood scores, and then applied statistical tests to

identify the AUs within each stratum that show significant discriminative power for MCI detection.

4. **Feature Fusion and MCI Classification:** Implement multimodal fusion for binary classification (MCI vs. Healthy).

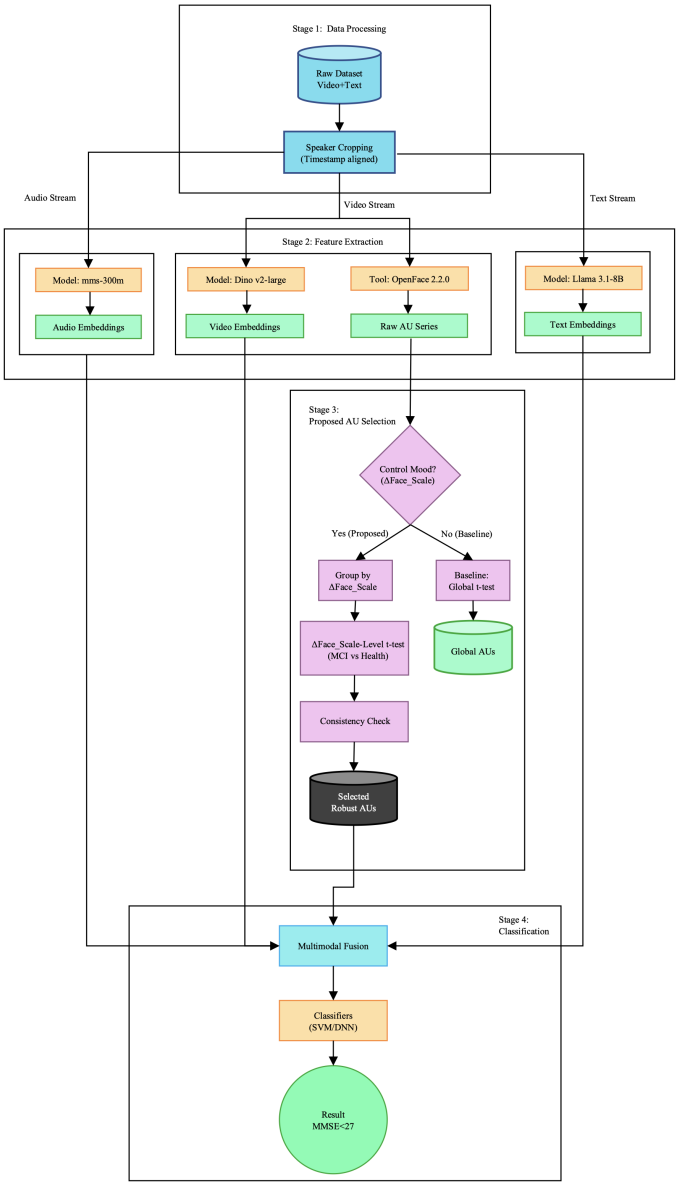

Figure 4.1: The architecture of this research

4.2 Phase separation

As mentioned in Introduction, the Coimagination Method consists of two phases: Introduction and Question-Answering (Q&A). In the Introduction phase, participants' interaction is dominated by passive listening or talking at the same time, leading to comparatively sparse behavioral cues for recognition. Consistent with the findings of Li et al., their evaluation indicated that MCI detection in this stage did not reach statistical significance. Accordingly, this study discusses and analyzes only the Q&A stage.

4.3 Data processing

To transform the continuous recordings into participant-centric analysis units, we implemented a straightforward video preprocessing pipeline that isolates the target participant in both time and visual space. First, in Time-based clipping, we aligned the continuous video stream with the transcription logs. For each utterance u, we used its start and end timestamps $\left(t_{\text{start}}^{(u)}, t_{\text{end}}^{(u)}\right)$ to extract a discrete clip from the original recording, ensuring that each clip strictly corresponds to the time window in which the target participant is speaking and remains synchronized with the matched audio and text segments. Second, in Region cropping, we reduced interference from the interviewer and background by applying a fixed-coordinate bounding box to each clip. Because the participant's position in the video feed is stable within a session, we predefined crop coordinates (x_1, y_1, x_2, y_2) and reused them across all clips from the same participant. The resulting clips serve as standardized visual inputs for downstream facial behavior analysis and video representation learning, while preserving utterance-level alignment across modalities.

4.4 Deep feature extraction

Following the spatio-temporal segmentation, the raw dataset consists of aligned triplets: cropped video clips, audio waveforms, and textual transcripts. To translate these heterogeneous signals into machine-interpretable representations capable of capturing the subtle biomarkers of MCI, we constructed a feature extraction pipeline utilizing three state-of-the-art pre-trained models.

4.4.1 Visual Representation (Dino v2-large)

To capture non-verbal facial cues, we employ Dino v2[36], a self-supervised Vision Transformer (ViT). Unlike traditional Convolutional Neural Networks (CNNs) trained on supervised classification tasks, which tend to focus on discriminative object classes, Dino v2 is trained using a self-supervised objective on a massive dataset of unlabelled images. This training paradigm allows the model to learn robust, object-centric features and fine-grained structural details without the bias of human annotations.

For the extraction process, we first sample frames from the cropped video clips at a fixed frame rate. Each frame is resized and normalized before being tokenized into patches and fed into the Dino v2 Transformer encoder. We specifically extract the output of the [CLS] token from the final transformer layer as the frame-level embedding. This [CLS] token serves as a global summary of the visual input, aggregating local patch information via the self-attention mechanism. Consequently, this representation effectively captures subtle facial micro-expressions and head pose variations that are critical for detecting the cognitive strain associated with MCI, providing a dense sequence of visual feature vectors for each utterance.

4.4.2 Paralinguistic Acoustic Modeling (mms-300m)

Speech pathology in MCI involves not just lexical content but also significant paralinguistic degradation, such as changes in speech rate, hesitation, and tone flatness. To analyze these acoustic properties, we utilize the mms-300m (Massively Multilingual Speech) model[34], which is built upon the wav2vec 2.0 architecture. Pre-trained on a vast corpus of speech data across over 1,000 languages, this model is highly sensitive to universal acoustic structures and phonetic nuances, making it superior to traditional handcrafted features like MFCCs(Mel-Frequency Cepstral Coefficients).

In our pipeline, the raw audio waveform corresponding to each utterance is input directly into the mms-300m model. The model processes the raw audio through a multi-layer convolutional feature encoder followed by a Transformer context network. We extract the latent representations from the final Transformer layer, which encode rich contextual information about the speaker's voice. These embeddings capture the minute variations in prosody and fluency that often precede clinical diagnosis, transforming the raw sound waves into high-dimensional vectors that reflect the participant's vocal stability and cognitive motor control.

4.4.3 Semantic Textual Analysis (Llama 3.1-8b)

Cognitive decline often manifests linguistically as a reduction in vocabulary richness and a loss of semantic coherence. To capture these high-level linguistic features, we leverage Llama 3.1[37], a state-of-the-art Large Language Model (LLM). While simpler embedding models (like Word2Vec)[38] focus on static word meanings, Llama 3.1 utilizes a deep Transformer-decoder architecture to model complex dependencies and long-range context within a sentence.

For each transcribed utterance, the text string is tokenized and processed by the Llama 3.1 network. Instead of using the model for text generation, we utilize it as a feature encoder by extracting the hidden state of the final token in the sequence from the last network layer. This specific embedding summarizes the entire semantic trajectory of the sentence, encapsulating both the syntactic structure and the logical coherence of the participant's speech. By leveraging the model's immense pre-trained knowledge base, we can detect subtle deviations in language production—such as repetitive speech or thematic drift—that serve as key textual biomarkers for MCI.

4.4.4 Facial Action Unit Extraction (OpenFace 2.2)

Complementing the deep semantic embeddings generated by Dino v2, we additionally employ the OpenFace 2.0 toolkit to extract interpretable, physically-grounded facial features. While Dino v2 captures global visual patterns, OpenFace is specifically utilized to detect the specific activation of Action Units (AUs) based on the Facial Action Coding System (FACS)[31].

For this study, we specifically utilize the classification output (AU_c) rather than the intensity regression (AU_r). The extraction process begins by aligning facial landmarks and analyzing localized HOG features. These features are processed through pre-trained Linear Kernel Support Vector Machines (SVMs) designed for binary classification. This approach yields a binary vector for each frame, indicating the presence (1) or absence (0) of specific facial muscle movements—such as AU4 (Brow Lowerer) or AU12 (Lip Corner Puller)—without the noise potentially introduced by fine-grained intensity estimation. By focusing on the definitive occurrence of these micro-expressions, we generate a robust, structured representation of facial dynamics. This binary data serves as the precise input for the subsequent "Mood-Controlled" statistical screening, allowing us to correlate the frequency of occurrence of specific AUs with mood and cognitive states.

AU	FACS name
AU01	Inner Brow Raiser
AU02	Outer Brow Raiser
AU04	Brow Lowerer
AU05	Upper Lid Raiser
AU06	Cheek Raiser
AU07	Lid Tightener
AU09	Nose Wrinkler
AU10	Upper Lip Raiser
AU11	Nasolabial Deepener
AU12	Lip Corner Puller
AU13	Cheek Puffer
AU14	Dimpler
AU15	Lip Corner Depressor
AU16	Lower Lip Depressor
AU17	Chin Raiser
AU18	Lip Puckerer
AU20	Lip Stretcher
AU22	Lip Funneler
AU23	Lip Tightener
AU24	Lip Pressor
AU25	Lips Part
AU26	Jaw Drop
AU27	Mouth Stretch
AU28	Lip Suck
AU45	Blink

Table 4.1: Common OpenFace Action Units (AUs) and their FACS names.

4.5 Mood-controlled AUs selection

4.5.1 Utterance-level AU activation

For each session s, we align the speech transcripts with the original video timeline, which yields a set of utterance segments. Let $\mathcal{F}_{s,u}$ denote the set of video frames belonging to the u-th utterance, with$N_{s,u} = |\mathcal{F}_{s,u}|$ frames. OpenFace outputs a binary activation signal $AU_{k,c}(f) \in \{0, 1\}$ for each frame, where k indexes the AU type and $f \in \mathcal{F}_{s,u}$ is the frame index. Here, the "frame index" simply refers to the sequential identifier of a frame within the utterance window; it is only used to enumerate all frames in $\mathcal{F}_{s,u}$.

We first compute the activation-frame ratio of AU k within the utterance:

$$r_{s,u,k} = \frac{1}{N_{s,u}} \sum_{f \in \mathcal{F}_{s,u}} AU_{k,c}(f)$$

This ratio measures how persistently an AU is activated during the utterance window: brief, sporadic activations lead to small $r_{s,u,k}$, whereas sustained activations lead to larger values.

We then introduce a ratio threshold $\tau \in \{0.1, 0.2, 0.3\}$ and define utterance-level AU activation as:

$$z_{s,u,k}(\tau) = \begin{cases} 1, & r_{s,u,k} > \tau, \\ 0, & r_{s,u,k} \leq \tau. \end{cases}$$

Here, $z_{s,u,k}(\tau) = 1$ indicates that utterance u is labeled as activating AU k under threshold τ, otherwise it is labeled as non-activated. This strategy mitigates over-sensitive decisions caused by single-frame fluctuations, requiring an AU to appear with a minimum proportion within the utterance window. In our experiments, we evaluate $\tau = 0.1/0.2/0.3$ to examine the impact of the threshold on feature stability and downstream classification performance.

4.5.2 Session-level AU Representation

For session-level modeling, we further aggregate utterance-level AU activations into a session-level representation. For each session s, let $\mathcal{U}_s$ be the set of utterances, with $M_s = |\mathcal{U}_s|$. We compute the session-level AU frequency (i.e., utterance-level occurrence rate) by averaging the binary utterance indicators:

$$x_{s,k}(\tau) = \frac{1}{M_s} \sum_{u \in \mathcal{U}_s} z_{s,u,k}(\tau), \qquad x_{s,k}(\tau) \in [0, 1]$$

Thus, $x_{s,k}(\tau)$ quantifies the proportion of utterances in session s that are labeled as activating AU k under threshold τ. Importantly, this representation does not directly average raw frame-level activations; instead, it first performs a robust utterance-level decision and then summarizes the prevalence of that behavior over the session, yielding an interpretable session-level AU feature vector for subsequent statistical screening and classification.

4.5.3 Δ Face Scale Stratification and Hypothesis Testing

To characterize session-level mood change, we define the Face Scale difference as:

$$\Delta F_s = F_s^{\text{after}} - F_s^{\text{before}}$$

Rather than collapsing ΔF_s into a small number of discrete categories, we treat each ΔF value as a stratum ℓ and define the set of sessions in that stratum as:

$$\ell = \Delta F_s, \qquad \mathcal{S}_\ell = \{s \mid \Delta F_s = \ell\}$$

This stratification aims to compare MCI and healthy controls under a comparable mood-change level, thereby reducing potential confounding effects induced by heterogeneous mood shifts across sessions.

Within each stratum ℓ, sessions are split into the MCI group and the Healthy group. For each AU feature $x_{s,k}(\tau)$, we perform an equal-variance independent-samples t-test (Student's t-test) to examine whether the group means differ significantly. The hypotheses are defined as:

$$H_0 : \ \mu^{(\ell)}_{\mathrm{MCI},k}(\tau) = \mu^{(\ell)}_{\mathrm{HC},k}(\tau), \qquad H_1 : \ \mu^{(\ell)}_{\mathrm{MCI},k}(\tau) \neq \mu^{(\ell)}_{\mathrm{HC},k}(\tau)$$

Let the sample sizes in stratum ℓ be $n^{(\ell)}_{\mathrm{MCI}}$ and $n^{(\ell)}_{\mathrm{HC}}$, with sample means/variances $\bar{x}^{(\ell)}_{\mathrm{MCI},k}, s^{2(\ell)}_{\mathrm{MCI},k}$ and $\bar{x}^{(\ell)}_{\mathrm{HC},k}, s^{2(\ell)}_{\mathrm{HC},k}$. Under the equal-variance assumption, we compute the pooled variance:

$$s^{2(\ell)}_{p,k}(\tau) = \frac{(n^{(\ell)}_{\mathrm{MCI}} - 1)s^{2(\ell)}_{\mathrm{MCI},k} + (n^{(\ell)}_{\mathrm{HC}} - 1)s^{2(\ell)}_{\mathrm{HC},k}}{n^{(\ell)}_{\mathrm{MCI}} + n^{(\ell)}_{\mathrm{HC}} - 2}$$

.

The corresponding t-statistic is:

$$t^{(\ell)}_k(\tau) = \frac{\bar{x}^{(\ell)}_{\mathrm{MCI},k} - \bar{x}^{(\ell)}_{\mathrm{HC},k}}{s^{(\ell)}_{p,k}(\tau)\sqrt{\frac{1}{n^{(\ell)}_{\mathrm{MCI}}} + \frac{1}{n^{(\ell)}_{\mathrm{HC}}}}}, \qquad \mathrm{df} = n^{(\ell)}_{\mathrm{MCI}} + n^{(\ell)}_{\mathrm{HC}} - 2$$

In practice, for each stratum ℓ and each threshold τ, we compute the t-statistic and the two-sided p-value for every AU dimension k. We then compare the p-value with a predefined significance level α (e.g., 0.05): if p$< \alpha$, we reject H_0 and regard AU k as statistically discriminative between MCI and healthy controls in that stratum under the given τ. The selected AUs are used as candidate features for downstream classification experiments, allowing us to evaluate the effectiveness of the stratified AU selection strategy.

4.6 Feature fusion and MCI classification

After deriving session-level representations, we conducted multimodal feature fusion and MCI classification. Let s denote a session. The session-level vectors from different modalities are denoted as: AU features $\mathbf{x}^{\mathrm{au}}_s \in \mathbb{R}^{d_{\mathrm{au}}}$,

audio embeddings $\mathbf{x}_s^{\text{aud}} \in \mathbb{R}^{d_{\text{aud}}}$, text embeddings $\mathbf{x}_s^{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$, and video embeddings $\mathbf{x}_s^{\text{vid}} \in \mathbb{R}^{d_{\text{vid}}}$. Audio/text/video embeddings were loaded from their corresponding .pt feature files. When a feature was provided as a temporal sequence (T, D), we applied mean pooling along the temporal dimension to obtain a fixed-length vector $(D,)$, enabling session-level alignment and fusion across modalities.

We systematically evaluated unimodal and multimodal settings. For any modality set $\mathcal{M} \subseteq \{\text{au}, \text{aud}, \text{txt}, \text{vid}\}$, we performed concatenation-based fusion at the session level:

$$\mathbf{x}_s^{(\mathcal{M})} = \text{concat}\Big(\{\mathbf{x}_s^m \mid m \in \mathcal{M}\}\Big),$$

where concat$(\cdot)$ denotes feature-wise concatenation to form a unified input vector for classification. Since AU descriptors and video embeddings are both derived from the visual channel and may carry overlapping information, we treated AU-based features and video embeddings as two alternative visual representation pathways. Accordingly, we evaluated AU-related combinations (e.g., $\{\text{au}, \text{aud}, \text{txt}\}$) and video-related combinations (e.g., $\{\text{vid}, \text{aud}, \text{txt}\}$) separately, without constructing a four-stream setting that simultaneously includes both AU and video, to preserve interpretability in comparisons.

We compared two classifiers: an SVM and a DNN. For the SVM, the fused vector $\mathbf{x}_s^{(\mathcal{M})}$ was standardized and then used to train a binary decision function, producing session-level predictions.

For the DNN model, we adopted a multi-branch architecture to learn modality-specific representations and perform feature fusion for MCI classification. For the i-th session sample, let $\mathbf{x}_i^m$ denote the available modality feature, where $m \in \{\text{AU}, \text{aud}, \text{txt}, \text{vid}\}$. Each modality is processed by an independent MLP branch $f_m(\cdot)$ that projects the raw feature into a compact latent representation:

$$\mathbf{z}_i^m = f_m(\mathbf{x}_i^m).$$

The outputs of the enabled branches are then concatenated to form a fused representation:

$$\mathbf{h}_i = \text{Concat}\big(\mathbf{z}_i^{m_1}, \mathbf{z}_i^{m_2}, \dots\big).$$

The fused vector $\mathbf{h}_i$ is fed into a classification head $g(\cdot)$ to produce a logit o_i, and the MCI probability is obtained via the sigmoid function:

$$o_i = g(\mathbf{h}_i), \qquad p_i = \sigma(o_i)$$

.

Here, $p_i \in [0, 1]$ is the predicted probability of MCI. We use binary labels $y_i \in \{0, 1\}$, where $y_i = 1$ indicates MCI and $y_i = 0$ indicates healthy controls. The model is trained using a class-weighted binary cross-entropy objective to address class imbalance:

$$\mathcal{L} = -w\, y_i \log p_i - (1 - y_i) \log(1 - p_i),$$

where w is the positive-class weight computed from the training set. To improve optimization stability and comparability across modalities, we standardize AU, audio, text, and video features independently. Regularization strategies, including dropout, weight decay, and early stopping based on validation loss, are applied to enhance generalization.

We evaluate the model using Leave-One-Group-Out (LOGO) cross-validation, where each group corresponds to all sessions from the same participant. In each fold, all sessions from one participant are held out for testing, and sessions from the remaining participants are used for training and validation. We report session-level Accuracy, Macro-F1, to assess subject-independent MCI detection performance and to quantify the contribution of different modality combinations.

4.7 Relationship framework among Face Scale, AUs, and MCI classification

The primary objective of this thesis is session-level mild cognitive impairment (MCI) classification from naturalistic online Coimagination conversations, where MCI labels are derived from clinical cognitive assessment (MMSE) using a threshold-based rule. In real-world interactive settings, however, behavioral signals are inherently multi-determined. Facial Action Units (AUs), together with speech and language modalities, may reflect cognitive-state differences, but they are also influenced by mood fluctuations, social strategies, and individual baselines. This confounding nature can reduce feature discriminability and weaken cross-subject generalization.

To explicitly characterize and mitigate mood-related confounding, this thesis incorporates Face Scale ratings collected before and after each session, using their difference as an observable proxy of mood change. Importantly, Face Scale is not treated as a supervision signal for MCI prediction. Instead, it serves as a mechanism to structure the analysis and guide robust AU selection. Within each mood-change level, statistical significance testing is conducted on AU-derived features to identify AUs that remain discriminative between MCI and healthy controls under comparable affective conditions. In

this pipeline, the statistical analysis module (t-test) functions as an evidence-driven filter that promotes robustness by reducing the influence of mood-driven variance on downstream learning.

In the machine learning stage, a linear-kernel SVM and a multi-branch DNN are implemented as the conventional baseline and the deep model, respectively. Systematic ablation studies are performed across unimodal and multimodal combinations. Both models are evaluated under the same Leave-One-Group-Out protocol with participant-level grouping, ensuring strict subject-wise separation between training and testing and emphasizing generalization to unseen participants. Overall, by integrating statistical testing with machine learning evaluation, this thesis establishes a unified framework linking Face Scale (mood change), AUs (facial behavior), and MCI (classification target), and uses it to interpret the impact of AU screening and multimodal fusion on classification performance.

Chapter 5

Experiments

5.1 Experimental setup

This chapter evaluates the proposed multimodal feature fusion framework for MCI binary classification. Experiments are conducted at the session level, where each sample corresponds to one participant's session with paired multimodal representations and a label. A participant-wise leave-one-out protocol is adopted to strictly separate training and testing by subject, enabling an unbiased assessment of generalization to unseen participants. Under the same splitting strategy and evaluation metrics, we compare different modality combinations, AU settings (filtered or unfiltered), and classifiers (SVM or DNN).

5.2 Dataset and label definition

The dataset involves 31 participants and 138 session-level samples. The original records are utterance-level (approximately 13,582 items), which are aggregated into session-level instances by grouping (participant, Session). Cognitive labels are derived from MMSE scores using a threshold of 27: sessions with MMSE < 27 are labeled as MCI, otherwise as Healthy. To maintain consistent semantics for binary outputs, we define MCI=1 and Healthy=0 in the DNN setting and use the same encoding for SVM for a fair comparison.

In addition to cognitive labels, each session is associated with two self-reported affective ratings (Face Scale) measured immediately before and after the session, denoted as F_s^{before} and F_s^{after}. The session-level mood change ΔF_s is computed and used for mood-change stratification and AU feature selection. Specifically, rather than collapsing mood change into three categories, we treat each distinct ΔF_s value as a stratum and conduct statistical signifi-

cance testing of AU features within each stratum, resulting in an AU subset under the "filtered AU" setting. Note that ΔF_s is only used for AU stratification/selection, while MCI/Healthy labels are determined solely by the MMSE threshold.

5.3 Feature fusion settings and significant AU selection experiments

This section discusses the experimental configurations for different modalities and their fusion combinations, and clarifies how statistically significant AU selection is incorporated into the cross-validation pipeline to ensure fair and consistent comparisons. In addition to OpenFace-based AU features, we include audio embeddings (1024 dimensions), text embeddings (4096 dimensions), and video embeddings (1024 dimensions). All unimodal and multimodal settings are evaluated under the same LOGO (Leave-One-Group-Out) protocol to enable a controlled comparison across fusion configurations.

For the feature fusion settings, we adopted a session-level early fusion framework. Audio, video, and transcripts were first aligned using utterance start/end timestamps so that all modalities correspond to the same semantic time window. We then extracted fixed-dimensional representations from pretrained encoders for audio, video, and text, while aggregating frame-level OpenFace outputs within each utterance segment to obtain interpretable utterance-level AU descriptors. To reduce sentence-level variability and measurement noise, all utterance representations belonging to the same participant and session were statistically aggregated (e.g., via mean pooling) to form stable session-level features. Finally, session-level vectors from the selected modalities were concatenated in a fixed order to form a unified representation that was fed into the classifier.

For the significant AU selection experiments, AU significance-based selection was performed strictly within each training fold to prevent information leakage. Specifically, we used LeaveOneGroupOut where each group corresponds to one participant and contains all of their sessions. In each fold, all sessions from the held-out participant formed the test set, while sessions from the remaining participants formed the training set. Within each fold, the complete feature processing pipeline—including AU selection, feature construction, and model training—was executed using training data only, and predictions were then generated for the corresponding held-out test sessions. Overall performance was computed by aggregating predictions across all folds, enabling leakage-free and robust evaluation of both fusion settings

and significant AU selection strategies.

We adopt LeaveOneGroupOut as the evaluation protocol, where each group corresponds to one participant containing all of their sessions. In each fold, the test set consists of all sessions from the held-out participant, while the remaining participants' sessions form the training set. To ensure a consistent and leakage-free evaluation, feature processing, AU selection, and model training are performed using training data only in each fold, and predictions are produced on the corresponding held-out test sessions. Overall metrics are obtained by aggregating predictions across all folds.

For AU features, we report results under two selection settings: ΔF_s-stratified AU selection and non-stratified AU selection. In both settings, independent two-sample t-tests are conducted within the training set to select AU dimensions that show significant differences between the MCI and Healthy groups. Specifically, for each fold and each proportion threshold $\tau \in \{0.1, 0.2, 0.3\}$, we first compute the session-level AU representation $x_{s,k}(\tau)$. We then perform equal-variance independent two-sample t-tests for each AU dimension using a two-sided test and a significance level α. In the stratified setting, training sessions are first stratified by ΔF_s, and significant AUs$\in \{AU01_c, AU10_c, AU17_c, AU45_c\}$ are selected within each stratum and aggregated; in the non-stratified setting, tests are conducted on the entire training set without stratification, and we can get AUs$\in \{AU10_c, AU23_c, AU25_c, AU45_c\}$. After determining the AU subset, both training and test AU features in that fold are restricted to the selected AU dimensions, which are then used alone or fused with other modality features for classification.

We systematically evaluate the following input settings: unimodal baselines include AU-only, audio-only, text-only, and video-only; bimodal settings include AU+audio, AU+text, audio+text, audio+video, and text+video; and trimodal settings include AU+audio+text and audio+text+video. Since AU and video features are both derived from video signals and may introduce overlapping or conflicting information in our data construction, we do not include AU+video fusion, and we also exclude the four-modality combination AU+video+audio+text to avoid redundancy and ambiguous interpretation. All feasible combinations are evaluated under the same LOGO protocol, using accuracy and macro F1 as primary metrics.

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

$$\text{Macro-F1} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i$$

Where "TP" is the number of true positives, "FN" is the number of false negatives, and "FP" is the number of false positives. "N" is the number of categories.

With these settings, the purpose of Section 4.3 is to provide a reproducible experimental framework to analyze: the effect of significant AU selection on AU-only performance and AU-based fusion, as well as the relative contribution of audio, text, and video features in AU-free multimodal combinations under a unified evaluation baseline.

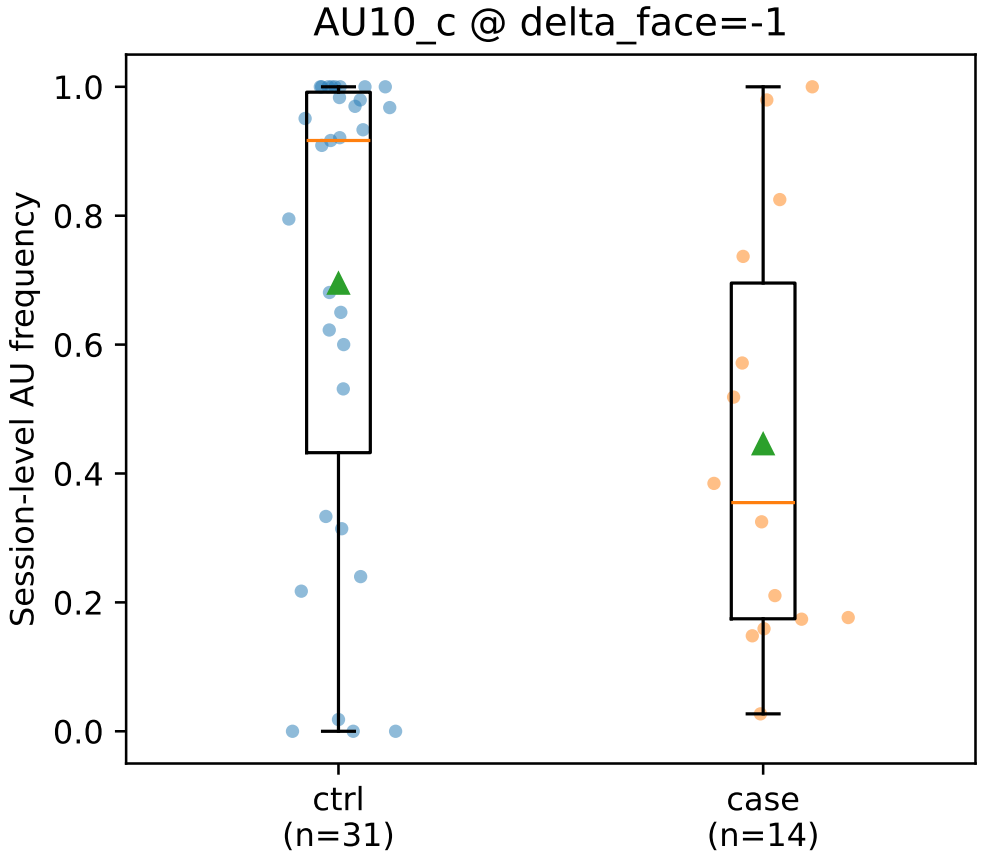

Figure 5.1: Boxplots of selected session-level AU features for the MCI(case) and Healthy(ctrl) groups. Each boxplot summarizes the distribution of the corresponding AU feature across session samples under a fixed frame-ratio threshold $\tau = 0.3$, where the box indicates the interquartile range and the center line denotes the median. This figure provides a visual comparison of group separability for the AUs retained after the significance-based selection.

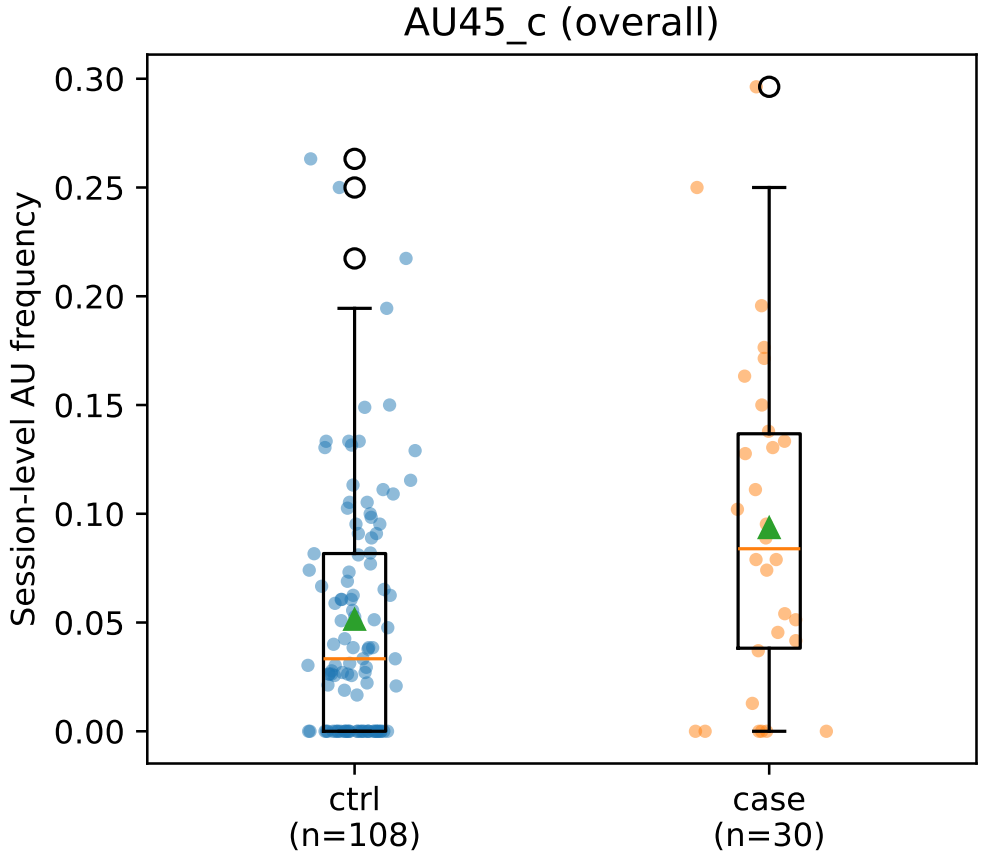

Figure 5.2: Boxplots of unselected session-level AU features for the MCI(case) and Healthy(ctrl) groups. Using the same visualization protocol as the selected-AU case, this figure reports the distribution comparison for the full AU set without feature selection, providing an overall view of group differences and a visual reference for subsequent classification-performance comparisons.

5.4 MCI prediction model

5.4.1 Support Vector Machine

This study first establishes a linear-kernel SVM as a conventional machine-learning baseline for binary classification on session-level samples. For each session, input features from different modalities are concatenated along the feature dimension to form a unified vector. We conduct ablation experiments

over both single-modality and multi-modality combinations to quantify the contribution of each modality to MCI discrimination. To prevent training bias caused by differences in feature scales, in each cross-validation fold we fit the standardization parameters using only the training set and apply the same transformation to the test set, ensuring a strict leakage-free evaluation protocol. Given class imbalance, the SVM is trained with balanced class weights to improve robustness to the minority class. When the text representation is high-dimensional, we apply Gaussian random projection to map it into a lower-dimensional space, thereby controlling feature size and improving training stability before concatenating it with other modalities.

For evaluation, we adopt Leave-One-Group-Out (LOGO) cross-validation, where each group corresponds to one participant; all sessions from the same participant are always assigned to the same side (training or testing). This design enables assessment of the model's generalization to unseen participants. We report session-level predictions and further aggregate session-level prediction scores at the participant level, producing participant-level results for overall performance reporting and statistical comparison.

Item	**Setting**
Classifier	Linear SVM (LinearSVC)
Kernel	Linear
C	1.0
Class weight	balanced
Max iterations	20000
Dual formulation	auto
Random state	0
Feature scaling	StandardScaler (fit on training only)
Cross-validation	Leave-One-Group-Out
Label threshold (MMSE)	27
Positive class mapping	case (MCI)=1 if MMSE$<$ 27
Text dimensionality reduction	GaussianRandomProjection
RP output dimension	1024
RP random state	0

Table 5.1: Hyperparameter settings for the linear SVM baseline.

5.4.2 Deep Neural Network

In the deep learning component, we adopt a multi-branch architecture to learn modality-specific representations independently and perform feature-level fusion at an intermediate layer. Specifically, each modality (AU, audio,

text, and video) is processed by an independent MLP branch that compresses the original high-dimensional embeddings into a more compact latent representation. The outputs of all branches are then concatenated along the feature dimension and fed into a classification head to produce the final prediction. The key motivation of this design is to preserve modality-specific information while enabling the network to capture cross-modal complementarity, thereby improving its ability to discriminate MCI. During training, we use fixed optimization and regularization settings (e.g., weight decay and dropout), and mitigate overfitting via validation-based monitoring and an early-stopping strategy, ensuring stable generalization even under a limited number of participants.

Consistent with the SVM baseline, the DNN is evaluated using Leave-One-Group-Out (LOGO) cross-validation, where each group corresponds to a participant, ensuring that both models follow the same data-splitting and evaluation protocol for fair comparison. The model outputs are saved at the session level and further aggregated at the participant level to obtain participant-level results for the final comparison. It is worth noting that we conducted systematic experiments across both unimodal and multimodal combinations. Moreover, since AU and video both represent facial-related information channels and may exhibit substantial redundancy or conflict, we avoided including AU and video simultaneously in multimodal fusion to reduce potential instability caused by overlapping representations.

Item	Setting
Architecture	Multi-branch MLP + concatenation fusion + MLP head
AU branch hidden dims	(32, 16)
Audio branch hidden dims	(256, 64)
Text branch hidden dims	(256, 64)
Video feature dimension	1024
Fusion	Concatenation (feature-level)
Classifier head	$64 \rightarrow 1$ (logit)
Activation	ReLU
Dropout	0.5
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	1×10^{-3}
Loss	BCEWithLogitsLoss
Class imbalance handling	pos_weight = $N_{\text{neg}}/N_{\text{pos}}$ (computed on training fold)
Epochs	80
Batch size	16
Early stopping patience	10 (based on validation loss)
Validation split	Participant-level split within training set
VAL_GROUP_FRAC	0.2
Cross-validation	Leave-One-Group-Out (group = participant)
Label threshold (MMSE)	27
Positive class mapping	MCI = 1
Random seed	0
Device	CUDA

Table 5.2: Hyperparameter settings for the multi-branch DNN model.

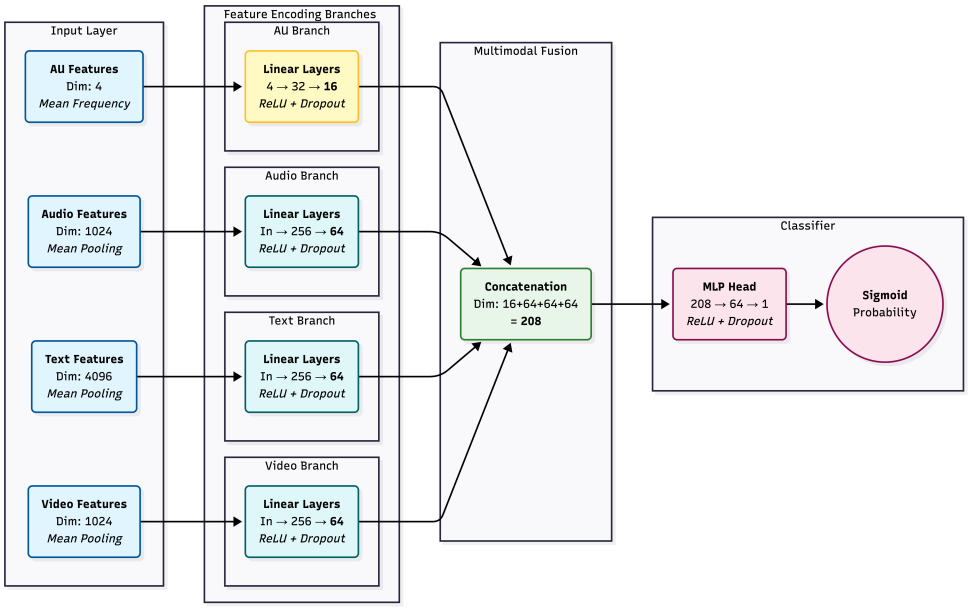

Figure 5.3: DNN

5.5 Results

Table 5.3 summarizes the session-level classification performance under settings where AU features are not included. Overall, the results vary noticeably across single modalities, and the relative strengths of modalities are not fully consistent between the two classifiers. In general, multimodal combinations tend to yield more stable performance than single-modality baselines, with audio–text related combinations showing particularly strong and consistent behavior. This observation suggests that, when facial behavior cues are excluded, linguistic and paralinguistic information still provides substantial discriminative signals for MCI classification. By contrast, combinations involving video exhibit larger fluctuations across models, indicating that the contribution of video features may be more dependent on the classifier and the fusion strategy. Collectively, these findings show that multimodal fusion can partially compensate for the limitations of individual modalities and improve robustness under the "no-AU" condition.

Modality setting	SVM (linear)		DNN	
	Acc	Macro F1	Acc	Macro F1
Audio-only	0.682	0.476	0.638	0.546
Text-only	0.615	0.511	0.664	0.624
Video-only	**0.698**	0.411	0.658	0.546
Audio + Text	0.635	0.532	**0.725**	**0.683**
Audio + Video	0.628	**0.586**	0.651	0.586
Text + Video	0.632	0.468	0.644	0.570
Audio + Text + Video	0.635	0.458	0.691	0.638

Table 5.3: Classification performance for modality combinations without AU features.

Table 5.4 reports results obtained using the full AU feature set, either alone or fused with other modalities. A key observation is that AU-only settings already provide meaningful discriminative capability, highlighting the usefulness of facial action representations for this task. However, when the full AU set is directly concatenated with other modalities, performance improvements are not consistently observed; in some settings, fusion leads to degradation or increased variability. This indicates that naïve feature-level fusion does not necessarily produce gains when the AU set is large, and the effectiveness of fusion can be influenced by redundancy, noise, and potential overlap among modalities. Therefore, these results motivate the subsequent comparison with screened AU features to assess whether reducing AU redundancy can facilitate more effective integration with other modalities.

Modality setting	SVM (linear)		DNN	
	Acc	Macro F1	Acc	Macro F1
AU-only	**0.681**	**0.618**	**0.742**	**0.675**
AU + Audio	0.616	0.455	0.604	0.549
AU + Text	0.674	0.537	0.685	0.648
AU + Audio + Text	0.580	0.473	0.671	0.629

Table 5.4: Classification performance using the full AU feature set without screening.

Table 5.5 presents the performance when AU features are screened via the proposed significance-based procedure. Compared with Table 5.5, the screened AU set yields more stable behavior for the linear baseline, particularly in AU-only and AU-plus-single-modality configurations, suggesting improved usability of the AU representation after screening. For the deep

model, the combination of screened AU and text features shows a clearer advantage, implying that removing less informative AU dimensions can make the complementarity between facial action cues and semantic information easier to exploit. Notably, not all AU-inclusive combinations improve after screening, indicating that AU screening primarily enhances the quality of AU representations, while the final fusion outcome still depends on how different modalities interact under a given model.

Modality setting	SVM (linear)		DNN	
	Acc	Macro F1	Acc	Macro F1
AU-only	**0.710**	0.642	0.624	0.579
AU + Audio	0.601	**0.645**	0.591	0.534
AU + Text	0.688	0.548	**0.711**	**0.681**
AU + Audio + Text	0.580	0.473	0.658	0.614

Table 5.5: Classification performance with significant AU features selected by the proposed screening procedure.

Taken together, Tables 5.3–5.5 support three main conclusions. First, without AU features, audio- and text-related modalities and their fusion provide relatively stable and competitive performance, confirming the value of linguistic and acoustic cues for MCI discrimination. Second, when AU features are introduced, the construction and selection of AU representations substantially affect downstream classification, and using the full AU set does not guarantee benefits under straightforward fusion. Third, screening AU features generally improves the practicality and stability of AU-based modeling—especially for the linear baseline—and can better reveal complementary effects with text in the deep model. Overall, these findings suggest that modality contributions are not simply additive; instead, careful feature screening and modality combination design are important for achieving robust performance.

Modality setting	Model	TN	FP	FN	TP
Audio-only	SVM	7	32	26	91
	DNN	81	30	24	14
Text-only	SVM	12	27	33	84
	DNN	74	37	13	25
Video-only	SVM	0	39	8	109
	DNN	86	25	26	12
Audio + Text	SVM	13	31	26	86
	DNN	81	30	11	27
Audio + Video	SVM	0	39	19	98
	DNN	78	33	19	19
Text + Video	SVM	6	33	24	93
	DNN	79	32	21	17
Audio + Text + Video	SVM	5	23	34	94
	DNN	80	31	15	23

Table 5.6: Confusion-matrix counts for modality combinations without AU features under LOGO evaluation. MCI is treated as the positive class (label=1).

Modality setting	Model	TN	FP	FN	TP
AU-only	SVM	75	33	11	19
	DNN	78	33	10	28
AU + Audio	SVM	80	28	25	5
	DNN	71	40	19	19
AU + Text	SVM	84	24	21	9
	DNN	75	36	11	27
AU + Audio + Text	SVM	71	37	21	9
	DNN	75	36	13	28

Table 5.7: Confusion-matrix counts for modality combinations using the full AU feature set (without screening) under LOGO evaluation. MCI is treated as the positive class (label=1).

Modality setting	**Model**	**TN**	**FP**	**FN**	**TP**
AU-only	SVM	79	29	11	19
	DNN	71	40	16	22
AU + Audio	SVM	78	30	25	5
	DNN	70	41	20	18
AU + Text	SVM	86	22	21	9
	DNN	74	37	14	24
AU + Audio + Text	SVM	71	37	21	9
	DNN	74	37	14	24

Table 5.8: Confusion-matrix counts for modality combinations using significant AU features selected by the proposed screening procedure under LOGO evaluation. MCI is treated as the positive class (label=1).

Chapter 6

Discussion

6.1 Performance without AU features

Table 5.3 reports the classification performance using modality settings that exclude AU features. Overall, the results indicate that the effectiveness of multimodal fusion depends on both the information content of each modality and the model's capacity to exploit cross-modal complementarity. In this setting, modalities such as audio, text, and video provide heterogeneous cues that relate to cognitive state from different perspectives (e.g., linguistic content, paralinguistic patterns, and visual behavioral dynamics). Consequently, performance varies across modality combinations, reflecting the extent to which the modalities contribute complementary rather than redundant information.

A consistent observation is that the deep model benefits more noticeably from feature fusion than the linear SVM baseline. This suggests that, when multiple high-dimensional embeddings are concatenated, a non-linear model with representation learning can better transform and align the heterogeneous inputs into a task-relevant space. In contrast, linear decision boundaries are more sensitive to noise dimensions and distribution shifts, which may limit the gains from naïve concatenation. Notably, combinations involving both semantic and acoustic channels tend to provide stronger evidence for MCI discrimination than unimodal settings, supporting the view that "what is said" (text) and "how it is said" (audio) are complementary cues under the session-level protocol.

6.2 Performance with the full AU feature set

Table 5.4 presents results when the full AU feature set is used without screening. In this condition, AU features demonstrate strong discriminative capability, particularly under the LOGO evaluation protocol that emphasizes generalization to unseen participants. This pattern implies that facial action unit statistics capture relatively stable behavioral signatures associated with cognitive status, and these signatures remain informative across individuals.

Importantly, adding audio and/or text features does not uniformly improve performance over AU-only. This does not imply that the additional modalities are uninformative; rather, it highlights a common challenge in multimodal learning under limited data: additional high-dimensional embeddings can introduce modality-specific nuisance variation (e.g., speaker identity, recording conditions, session-dependent factors) that competes with the cognitive signal. When such variation is not sufficiently suppressed or disentangled, fusion may yield limited gains or even degrade robustness. Under participant-independent testing, this effect can become more pronounced because participant-specific cues are not transferable to the held-out subject.

6.3 Performance with significant AU features selected by screening

Table 5.5 summarizes results using AU features selected by the proposed screening procedure based on Face Scale stratification and within-stratum significance testing. The results reveal an algorithm-dependent effect of AU screening. For the linear SVM, screening generally strengthens the AU-only setting, which is consistent with the role of feature selection in enhancing linear separability by removing weak or noisy dimensions. By constraining the input to a compact subset of AUs that exhibit significant group differences within Face Scale strata, the SVM can fit a more stable hyperplane with reduced susceptibility to irrelevant variation.

In contrast, the deep model does not necessarily achieve its best performance with screened AU-only features. This observation is consistent with the fact that deep networks often benefit from richer joint patterns and higher-order interactions among features, rather than relying solely on a small set of individually significant dimensions. After screening, AU features may become more specialized and condition-dependent (i.e., informative primarily within certain Face Scale strata), which can reduce their standalone consistency across the entire dataset. In this case, the screened AU features appear to serve more effectively as complementary cues when fused with

semantically stable modalities such as text, enabling the deep model to exploit cross-modal synergy. This explains why AU+Text becomes particularly competitive for the DNN under the screened setting, where AU contributes additional behavioral evidence while text provides a strong, more globally consistent signal.

6.4 Cross-table summary

Taken together, Tables 5.3-5.5 support three main conclusions. First, in the absence of AU features, deep multimodal fusion is generally more effective than a linear baseline, indicating the value of non-linear representation learning for heterogeneous high-dimensional embeddings.

Second, when using the full AU feature set, AU-only can already provide strong and robust discrimination under participant-independent evaluation, while adding other modalities may introduce nuisance variation that limits fusion gains.

Third, the Face Scale–based AU screening procedure improves the effectiveness of AU features for linear classification, whereas in the deep model the screened AUs function most reliably as complementary signals when fused with text, rather than as a standalone modality. These findings highlight that the benefit of AU screening is model-dependent and that effective multimodal fusion should consider both feature reliability under distribution shift and the model's ability to learn cross-modal interactions.

Chapter 7

Conclusion

This thesis studied session-level MCI classification from online Coimagination meetings using multimodal behavioral signals. Based on a Japanese dataset containing 31 participants and 40 recorded meeting videos, we constructed aligned audio, video, and transcript segments at the utterance level and derived session-level representations for modeling. MCI labels were defined using the MMSE threshold of 27, and mood information was provided by Face Scale ratings (7 levels from 6 to 0) recorded before and after each session, enabling session-wise mood change to be quantified by ΔFace Scale. Under this setting, the overall goal of this work was to detect cognitive status from natural group conversations while mitigating confounding effects introduced by affective variation.

Methodologically, the main contribution is a mood-controlled AU feature selection strategy designed to reduce the ambiguity of facial expressions. Because facial behaviors may correspond to either cognitive difficulty or temporary mood states, we used ΔFace Scale to stratify sessions and conducted within-stratum significance testing to select AU features that show discriminative differences between MCI and healthy groups. This approach provides a principled way to screen facial features before multimodal learning, and it directly targets the confounding between affective changes and cognitive signals. For classification, two representative paradigms were implemented: a linear SVM baseline with standardized inputs and a multi-branch DNN that performs modality-specific representation learning followed by feature-level fusion for prediction.

Empirical results indicate that different modalities contribute unequally to MCI discrimination and that their effectiveness depends on the learning model. When excluding AU information, Audio+Text yields the most reliable performance for the DNN, suggesting that linguistic content and acoustic patterns provide complementary cues in conversational assessments.

When incorporating AU information, AU-related features show strong discriminative potential in the full feature configuration, and the proposed AU screening procedure can improve stability and performance under specific settings, especially for linear models. Overall, the results support the conclusion that multimodal fusion is beneficial for conversational MCI detection, but facial behavioral cues require careful treatment with explicit control of affect-related variability to avoid unstable or misleading patterns.

Despite these contributions, several limitations remain. First, the dataset size and the number of participants are limited, which constrains statistical power and may amplify individual variability. Second, Face Scale provides only coarse session-level mood annotations and may not fully capture intra-session affect dynamics. Third, some modality pairs (e.g., AU and face-derived video representations) may contain redundancy, and more robust representation learning strategies may be needed to exploit them jointly. Future work will focus on expanding the dataset, incorporating temporal modeling to capture within-session dynamics, and developing methods that explicitly disentangle cognitive markers from affective signals. In addition, more systematic interpretability analyses that link multimodal patterns to clinically meaningful behaviors could further strengthen the practical value of the proposed framework.

Chapter 8

Acknowledgments

I would like to express my sincere gratitude to everyone who has supported and guided me throughout my study and research. First and foremost, I am deeply grateful to my supervisor, Prof.Okada. It was under his guidance that I was introduced to the field of multimodal machine learning. He has taught me essential research skills and, more importantly, the philosophy of conducting scientific research. His rigorous attitude, insightful advice, and continuous encouragement have been fundamental to the completion of this work.

I would also like to thank Assistant Professor Li Sixia for his invaluable support and inspiration. He introduced me to a wide range of practical research methods, encouraged me to examine problems from different perspectives, and provided many insightful suggestions that greatly enriched my thinking and motivated my progress.

My gratitude also goes to JAIST for providing an excellent research environment and a broad academic platform that enabled me to learn and grow. In addition, I would like to thank my parents for their unwavering support and encouragement throughout my journey. Finally, I am grateful to all members of my laboratory for their help in both research and daily life. I truly enjoyed the time we spent discussing research topics together, and those moments have been an important and memorable part of my experience.

Bibliography

[1] World Health Organization. ,"world failing to address dementia challenge", available at: https://www.who.int/news/item/02-09-2021-world-failing-to-address-dementia-challenge.

[2] C. P. Guruge et al. Advances in multimodal behavioral analytics for early dementia diagnosis: A review. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*, pages 328–340, 2021.

[3] M. Otake et al. Coimagination method: Communication support system with collected images and its evaluation via memory task. In *International Conference on Universal Access in Human-Computer Interaction*, pages 403–411, 2009.

[4] S. Li et al. Automatic mild cognitive impairment estimation from the group conversation of coimagination method. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24)*, pages 355–360, 2024.

[5] Ying Zhou, Xiuyu Yao, Wei Han, Yingxin Li, Jiajun Xue, and Zheng Li. Measurement of neuropsychiatric symptoms in the older adults with mild cognitive impairment based on speech and facial expressions: a cross-sectional observational study. *Aging & Mental Health*, 28(5):828–837, 2024.

[6] I. Arevalo-Rodriguez et al. Mini-mental state examination (mmse) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (mci). *Cochrane Database of Systematic Reviews*, (3):CD010783, 2015.

[7] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[8] Susanne G. Mueller, Norbert Schuff, Kristine Yaffe, Catherine Madison, Bruce Miller, and Michael W. Weiner. Hippocampal atrophy patterns in mild cogni-

tive impairment and alzheimer's disease. *Human Brain Mapping*, 31(9):1339–1347, 2010.

[9] Adam S. Fleisher, Kewei Chen, Xiaofen Liu, Auttawut Roontiva, Pradeep Thiyyagura, Napatkamon Ayutyanont, Abhinay D. Joshi, Christopher M. Clark, Mark A. Mintun, Michael J. Pontecorvo, P. Murali Doraiswamy, Keith A. Johnson, Daniel M. Skovronsky, and Eric M. Reiman. Using positron emission tomography and florbetapir f18 to image cortical amyloid in patients with mild cognitive impairment or dementia due to alzheimer disease. *Archives of Neurology*, 68(11):1404–1411, 2011.

[10] Serge A. R. B. Rombouts, Frederik Barkhof, Rutger Goekoop, Cornelis J. Stam, and Philip Scheltens. Altered resting state networks in mild cognitive impairment and mild alzheimer's disease: an fmri study. *Human Brain Mapping*, 26(4):231–239, 2005.

[11] M. Asgari et al. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228, February 2017.

[12] J. Tröger et al. Telephone-based dementia screening i: Automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '18)*, pages 59–66, 2018.

[13] A. others König. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, March 2015.

[14] L. Tóth et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138, 2018.

[15] B. Mirheidari et al. Diagnosing people with dementia using automatic conversation analysis. In *Interspeech 2016*, pages 1220–1224, 2016.

[16] K. Lopez-de Ipina et al. Advances on automatic speech analysis for early detection of alzheimer disease: A non-linear multi-task approach. *Current Alzheimer Research*, 15(2):139–148, 2018.

[17] J. Tröger et al. Automated speech-based screening for alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '17)*, pages 292–297, 2017.

[18] K. C. Fraser, N. Linz, et al. Predicting MCI status from multimodal language data using cascaded classifiers. *Frontiers in Aging Neuroscience*, 11:205, August 2019.

[19] K. López-de Ipiña, J. B. Alonso, et al. On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7(1):44–55, 2015.

[20] J. J. G. Meilán et al. Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334, 2014.

[21] K. Horley et al. Emotional prosody perception and production in dementia of the alzheimer's type. *Journal of Speech, Language, and Hearing Research*, 53(5):1132–1146, October 2010.

[22] K. E. Forbes-McKay et al. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological Sciences*, 26:243–254, 2005.

[23] T. Endo et al. Initial response time measurement in eye movement for dementia screening test. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 262–265, 2017.

[24] A. Oyama et al. Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. *Scientific Reports*, 9:12932, 2019.

[25] T. D. W. Wilcockson et al. Abnormalities of saccadic eye movements in dementia due to alzheimer's disease and mild cognitive impairment. *Aging*, 11(15):5389–5398, 2019.

[26] D. Lagun et al. Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of Neuroscience Methods*, 201(1):196–203, 2011.

[27] M. Pasqualini et al. Facial expression in mild dementia of the alzheimer type. *Behavioural Neurology*, 8:149–156, January 1995.

[28] K. Asplund et al. Facial expressions in severely demented patients—a stimulus–response study of four patients with dementia of the alzheimer type. *International Journal of Geriatric Psychiatry*, 6(8):599–606, 1991.

[29] S. Lautenbacher et al. Facial pain expression in dementia: A review of the experimental and clinical evidence. *Current Alzheimer Research*, 14(5):501–505, 2017.

[30] M. Kunz et al. The facial expression of pain in patients with dementia. *Pain*, 133(1-3):221–228, 2007.

[31] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[32] T. Baltrušaitis et al. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[33] A. Asgarian et al. Limitations and biases in facial landmark detection: An empirical study on older adults with dementia. *arXiv*, May 2019.

[34] B. Taati et al. Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia. *IEEE Access*, 7:25527–25534, 2019.

[35] G. Gosztolya et al. Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, 53:181–197, 2018.

[36] M. Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023.

[37] A. Grattafiori et al. The llama 3 herd of models. *arXiv*, 2024.

[38] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.