Multimodal Classification of Alzheimer's Disease by Combining Facial and Eye-Tracking Data

Shih-Han Chou SHCHOU75@CS.UBC.CA
Miini Teng MIINI.TENG@UBC.CA
Harshinee Sriram HSRIRAM@CS.UBC.CA
Chuyuan Li CHUYUAN.LI@UBC.CA
Giuseppe Carenini CARENINI@CS.UBC.CA
Cristina Conati CONATI@CS.UBC.CA
Thalia S. Field THALIA.FIELD@UBC.CA
University of British Columbia, Canada

Hyeju Jang HYEJUJ@IU.EDU
Indiana University-Purdue University Indianapolis, Indianapolis, USA

Gabriel Murray GABRIELM@CS.UBC.CA
University of British Columbia, Canada

Abstract

In recent years, there has been growing interest in developing a non-invasive tool for detecting Alzheimer's Disease (AD). Previous studies have shown that a single modality such as speech or eye-tracking (ET) data can be effective for classifying AD patients from healthy individuals. However, understanding the role of other modalities, and especially the integration of facial analysis with ET for enhancing dementia classification, remains under-explored. In this paper, we investigate whether we can leverage facial patterns in AD patients by building on EMOTION-FAN—a deep learning model initially developed for recognizing seven distinct human emotions, now fine-tuned for our facial analysis tasks. We also explore the efficacy of leveraging multimodal information by combining the results from the facial and ET data through a late fusion technique. Specifically, our approach uses a neural classifier to learn from raw ET data (VTNet) alongside the fine-tuned EMOTION-FAN model that learns from the facial data. Experimental results show that facial data gives superior results than ET data. Notably, we obtain higher scores when both modalities are combined, providing strong evidence that integrating multimodal data benefits performance on this task.

Keywords: Alzheimer's disease, cognitive impairment, machine learning, multimodal, video, eye-tracking, facial patterns

Data and Code Availability The dataset we use for the experiments is from Jang et al. (2021). The dataset is not available to researchers due to patient privacy. Our code is available at https://github.com/ShihHanChou/AD_facial_ET/.

Institutional Review Board (IRB) The dataset we use for this paper comes from the CANARY project at University of British Columbia (UBC), which was approved by the UBC clinical research ethics board (H17-02803-A036).

1. Introduction

Dementia refers to cognitive impairment that affects independent activities of living (Revi, 2020). It is associated with a number of neurodegenrative conditions, with Alzheimer's Disease (AD) being most prevalent (Scharre, 2019). Currently, more than 55 million people have dementia worldwide, and every year, there are nearly 10 million new cases (WHO). AD typically unfolds in three stages—early, middle, and late—also known as mild, moderate, and severe. A successful disease-modifying drug would be most likely to demonstrate an effect in individuals who do not yet have advanced neurodegenerative changes (Reiman et al., 2016), which is why early detection of AD is of great importance. The detection of dementia typically relies on traditional methods such as cognitive tests, clinical evaluations, and brain scans. Although these methods are effective, they are resource-intensive, costly,

and often invasive, which makes widespread application difficult. Therefore, there is a critical need for more accessible, efficient, and less invasive methods for detecting subtle indicators of dementia.

Machine learning models, built on non-invasive patient data, are potentially good candidates to be used as non-invasive screening tools. Previous work has explored feature extraction strategies and different machine learning methods, such as Random Forest, Logistic Regression, and Neural Networks (Kong et al., 2019; Pulido et al., 2020; Sheng et al., 2022) to classify AD patients and healthy controls, mostly with speech (Fraser et al., 2016) and eye-tracking data (Sriram et al., 2023). More recent work has investigated the added value of combined speech and eye-tracking modalities (Fraser et al., 2019; Barral et al., 2020; Jang et al., 2021). However, the role of other modalities, such as video information containing facial expressions, remains under-investigated.

On the other hand, some clinical studies have shown association between facial expressions and dementia. For instance, Cai et al. (2021); Asplund et al. (1991) found that AD is correlated with apathy and decreased facial expressivity. It was also observed that, in certain cultural environments, people with AD tend to smile more (Matsushita et al., 2018). Chien et al. (2023) investigated the facial asymmetry between AD patients and healthy controls and showed that AD patients exhibited increased facial asymmetry. Despite the promising clinical evidence, facial data has not been well explored for AD classification. Further, no research has investigated the potential benefits of combining facial data with other modalities.

In this paper, we address this gap by investigating the use of facial videos to distinguish between AD and healthy controls. Given that videos capture facial patterns and may offer temporal information, we hypothesize that this information is valuable to distinguish AD patients from healthy controls. Specifically, we aim to leverage the differences in facial patterns from an emotion-based model, i.e., the EMOTION-FAN framework (Meng et al., 2019), originally designed for facial emotion recognition. In practice, we compare two strategies to explore the EMOTION-FAN framework to analyze facial patterns relevant to dementia detection: linear modeling vs. fine-tuning. Our experiments were conducted on a multimodal dataset (Jang et al., 2021) in which videos were collected for four tasks: Pupil Calibration, Picture Description, Reading, and Memory Description. On all these tasks, we found that facial data delivered remarkable results (AUC $>$ 80%). Additionally, we explored the efficacy of multimodal inference by combining facial and eye-tracking (ET) data. For this purpose, we used a simple yet effective late fusion strategy, namely the average voting mechanism. This fusion approach greatly improved the performance of a single ET model or a facial model, demonstrating the advantages of multimodal fusion in AD classification tasks.

Our contributions in this paper are threefold:

- We explore a new modality for AD classification by using facial videos. To the best of our knowledge, we are the first to apply deep learning methods using facial patterns for this task.
- We propose a simple yet effective fine-tuning based on the EMOTION-FAN model to capture differences in facial patterns between AD/control.
- We significantly improve AD classification results by combining facial and ET data. We show the complementary effects of different modalities through a comprehensive error analysis.

2. Related Work

2.1. Facial and Eye-Tracking Information for AD Classification

Clinicians have examined differences in facial expression patterns in patients with Alzheimer's Disease (AD) to explore clinical significance and disease correlation (Seidl et al., 2012). Research has shown a connection between Alzheimer's disease (AD) and apathy, leading to reduced facial expressivity (Cai et al., 2021; Asplund et al., 1991). This decrease in expressivity aligns with observations that patients in advanced stages of AD frequently face difficulties in displaying facial emotional responses, signaling a decline in facial responsiveness (Asplund et al., 1991). There has also been evidence that lower cognitive screening scores was correlated with diminished intensity and variability of smiles (Jonell et al., 2021). Conversely, a study observed that individuals with dementia might display increased facial expressiveness due to reduced control over facial expressions (Smith, 1995). Facial expressions were also different across diverse cultures, where patients with AD were seen to smile more in Asian cultures to "save face", further underscoring the complex relationship between cognitive function and facial expression patterns (Matsushita et al., 2018). Chien et al. (2023) studied facial symmetry—in the areas of eyes, mouth, and eyebrows—of individuals with and without cognitive impairment. They hy-

pothesized that since facial asymmetry would increase with age, this could be a potential AD marker.

There has been recent interest in using facial data and AI for AD classification. For instance, Umeda-Kameyama et al. (2021) attempted to differentiate pictures of 121 patients with cognitive impairment and 117 cognitively sound participants, using simple CNNs and various deep learning models such as Xception (Chollet, 2017), SENet50 (Hu et al., 2018), ResNet50 (He et al., 2015), and VGG16 (Simonyan and Zisserman, 2014). They found that the Xception (Chollet, 2017) model was most promising in differentiating still facial images of dementia patients from controls, achieving promising sensitivity (87.31%), specificity (94.57%), and accuracy (92.56%).

Another recent study by Zheng et al. (2023) explored using face-related features from video data to detect dementia. They employed several classification algorithms such as LSTM (Hochreiter and Schmidhuber, 1997), ResNet (He et al., 2016b), and VGG16 (Simonyan and Zisserman, 2014) and found that the HOG features obtained an accuracy of 79%. These innovative approaches highlighted the potential of AI in providing non-invasive, effective tools for early detection and monitoring of cognitive impairment through facial analysis.

Research on eye-tracking (ET) data as an informative source for AD detection has been spurred by findings that AD impacts eye functionality, leading to notable differences in eye movements such as fixations, saccades, and pupil responses (Garbutt et al., 2008; Molitor et al., 2015; MacAskill and Anderson, 2016).

Prior studies have trained classifiers on ET data features selected based on their relevance to visual tasks known to be affected by AD. For example, Pavisic et al. (2017) developed classifiers using features critical for evaluating task performance involving fixation stability and the tracking of stimuli. Biondi et al. (2017) used features pertinent to Reading tasks, such as the frequency of repeated fixations on words. Jang et al. (2021) focused on features from tasks including Pupil Calibration and Reading, emphasizing regions of interest within these tasks, such as specific areas of a picture or text. These studies generally achieved accuracy rates over 80%, although the datasets in studies (Pavisic et al., 2017) and (Biondi et al., 2017) were relatively small, with 57 and 69 data respectively, and they only reported classification accuracy. In contrast, Jang et al. (2021) used a larger dataset of 126 data and conducted a comprehensive analysis using multiple performance metrics. Hence, we used the dataset from Jang et al. (2021) for our experiments.

2.2. Multi-modalities for AD Classification

Speech and eye-tracking modalities have been explored to characterize dementia, highlighting the potential use of these features to augment screening tools (Jang et al., 2021; Fraser et al., 2016; Sheng et al., 2022; Barral et al., 2020). Barral et al. (2020) and Jang et al. (2021) used a late fusion technique to combine predictions from different models based on ET and speech analysis, to enhance the overall predictive accuracy for AD classification. In their experiments, data from different modalities were first used to train individual classifiers such as logistic regression, random forest, and Gaussian Naïve Bayes. These individual predictions were then aggregated using an averaging method, where the final classification decision was derived by calculating the mean of the prediction probabilities from each model. Language-only and eye-movement-only models achieved an AUC of 0.73 and 0.77, respectively. However, a late fusion approach combining multimodal language and eye movement data significantly increased the overall performance to 0.80 (Barral et al., 2020). In another study, Zheng et al. (2023) also used speech and ET data, but with different feature extraction techniques. They used features extracted from the *bottleneck* layer—a layer that has fewer units than other layers—in a deep neural network (DNN). This architecture forced the network to compress information passing through it, which helped to extract the most salient features for the task at hand. Both papers demonstrated multi-modal approach as more effective for dementia detection than single-modality models.

Existing research on the screening of AD using facial patterns has predominantly focused on still images (Matsushita et al., 2018; Chien et al., 2023). In contrast, our study employs video data to enhance the dynamic analysis of facial patterns. Despite facial data having the potential to augment AD detection (Umeda-Kameyama et al., 2021; Zheng et al., 2023; Chien et al., 2023), there is a lack of research exploring the combination of facial and ET data, which our study aims to address.

3. Dataset

We use the dataset from Jang et al. (2021) for this study. This dataset comprises data gathered from pa-

tients at a specialized memory clinic, either diagnosed with Alzheimer's Disease (AD) or exhibiting initial symptoms of mild cognitive impairments potentially progressing to AD. Additionally, control participants from the community were included, and matched with the patient group based on sex and age. See Appendix A for participant demographics.

Participants were positioned at a testing platform equipped with a Tobii-Pro X3-120 eye tracker (120 Hz sampling rate) at the screen's base for gaze coordinates, head distance, and pupil data tracking. Videos of participants were captured using Logitech C922x Pro Stream Webcam 1080P Camera. Participants completed four tasks: a Pupil Calibration task, a Picture Description task, a paragraph Reading task, and a memory recall task devoid of visual elements. Comprehensive study details are accessible in Jang et al. (2021). We leverage facial and ET data collected during the initial four tasks:

- **Pupil Calibration**. Participants fixated on a static target for 10–15 seconds to detect square-wave jerks, indicative of AD.
- **Picture Description**. Participants verbally described the Cookie Theft picture from the Boston Diagnostic Aphasia Examination , a widely used task for evaluating spontaneous speech in various clinical contexts , including AD.
- **Reading**. Participants read aloud a standardized paragraph from the International Reading Speed Texts (IReST). This passage comprises 155 words and discusses how flora and fauna adapt in arid environments across 9 sentences. This task aimed to identify common reading deficits in AD, such as diminished reading speed and increased instances of word fixations or re-fixations.
- **Memory**. Participants describe a pleasant past experience to capture additional spontaneous speech data. The goal is to elicit speech deficits that may be missed in the Picture description or reading task. Additionally, the lack of visual stimuli allows the task to be completed identically despite possible variation in participant vision (e.g., low visual acuity, or blurred vision).

Completing these tasks took an average of 7 minutes. In total, this dataset contains 144 participants where 75 are control patients (average age = 72, standard deviation = 9) and 69 are AD patients (average age = 69, standard deviation = 15).

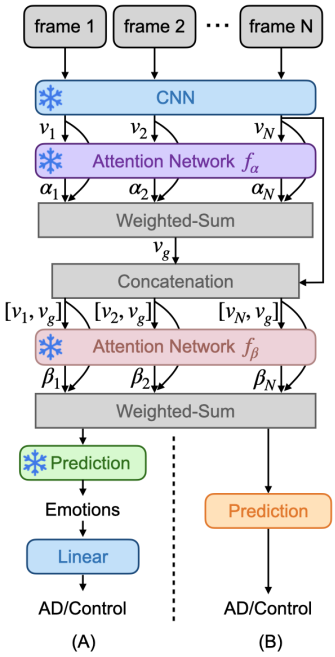

Figure 1: Model Overview. (A) EMOTION-FAN+Linear Model. (B) Fine-tuned EMOTION-FAN Model (ours). Frozen modules are marked with snowflakes.

4. AD Classification Using Facial Data

To explore facial patterns from videos that might be indicative of AD, we investigate potential differences in facial emotions between AD patients and healthy controls. Facial expressions associated with emotions are readily detectable and interpretable features. Specifically, we focus on the facial emotions including *happy, angry, disgust, fear, sad, surprise, neutral* (Matsushita et al., 2018; Cai et al., 2021). To this end, we first leverage an off-the-shelf emotion recognition model, EMOTION-FAN (Meng et al., 2019), which is pre-trained on the AFEW dataset (Dhall et al., 2018) to generate emotion probabilities of the input facial videos. In the second step, we use features extracted before the final emotion prediction step and fine-tune the last linear layer for AD classification.

4.1. Emotion-based Model Architecture

We use the facial recognition model, called EMOTION-FAN (Meng et al., 2019), a deep learning model initially developed for recognizing seven distinct human emotions. This model (see Figure 1) comprises a CNN network to extract facial representations from images, two attention networks (f_α, f_β) to learn frame and

video features, respectively, and a prediction layer for the emotion prediction.

When EMOTION-FAN is given a sequence of N image frames as input, ResNet18 (He et al., 2016a) (i.e., the CNN network) takes each frame i and generates a representation v_i. The image representations v_i are then passed into the first attention layer f_α to obtain local self-attention weights α_i. Precisely, f_α contains linear layers and a Sigmoid function $\sigma(\cdot)$. α_i is defined as follows:

$$\alpha_i = \sigma(f_\alpha(v_i)) \tag{1}$$

EMOTION-FAN also uses a global attention weight v_g which is obtained by aggregating the local weights α_i from the frame representations. v_g is simply the weighted sum of all α_i:

$$v_g = \sum(\alpha_i) \tag{2}$$

The local and the global attention are concatenated to form a final relation-attention weight β_i for each frame i, which is calculated using an additional linear layer (f_β) and a Sigmoid function $\sigma(\cdot)$:

$$\beta_i = \sigma(f_\beta([v_i, v_g]^T)) \tag{3}$$

Originally, the EMOTION-FAN model uses the final representation β_i as the input of the classification layer for emotion recognition. Here, we use EMOTION-FAN as the backbone and introduce two variations for our AD/health classification task:

(A) EMOTION-FAN+Linear Model. In this variation, an additional linear layer (shown in blue in Figure 1 (A)) is added after the emotion prediction layer, directly transforming the predicted emotion probabilities into a binary class prediction of AD patients and healthy controls. In this setup, we explore whether it is feasible to directly detect AD patients from facial emotions. During training, all parameters in the EMOTION-FAN backbone are frozen. We only fine-tune the newly added linear layer.

(B) Fine-tuned EMOTION-FAN Model (ours). In this variation, we leverage the features from the EMOTION-FAN model before the prediction layers. The objective of this is to leverage facial embeddings obtained in the emotion recognition task for AD classification. Specifically, we fine-tune the last linear layer in EMOTION-FAN while freezing previous relation-attention layers during training. As shown in Figure 1 (B), we freeze the CNN, Attention Network f_α, and Attention Network f_β. The fine-tuned part is the Prediction layers. We re-trained the model using cross-entropy loss.

4.2. Experimental Setup

Video Preprocessing. The EMOTION-FAN model takes a sequence of face-centered images as input. To obtain images from our video data, we first preprocess the videos following the process in Meng et al. (2019). Precisely, we convert the videos into a sequence of still images, capturing one frame per second (FPS). On average, we have 32, 105, 110, and 88 frames for the Pupil Calibration, Picture Description, Reading, and Memory tasks, respectively. The image frames are then processed using the Dlib toolbox (King, 2009) for face detection and alignment. The bounding boxes around the faces are expanded by 25% to ensure that the entire face of the participant was captured. The resulting face images are then cropped and resized to 224×224 resolution. See Appendix B for more details.

Implementation Details. We use the PyTorch framework for model implementation. We adopt the pretrained EMOTION-FAN model from the original paper (Meng et al., 2019) and fine-tune it by modifying the final classification layers. Following the hyper-parameter settings in Meng et al. (2019)'s work, the batch size for loading image frames is set to be 48 for training and 65 for validation and testing. We employ the SGD optimizer with an initial learning rate of $4e^{-3}$, a momentum of 0.9, and a weight decay of $1e-4$. Each model is trained for 10 epochs without early stopping. We use the validation set to find the best checking point. All the experiments are performed on a single GPU with 12GB memory on a single server. It takes about 0.7 days training for the Pupil Calibration task, 2 days for the Picture Description task, and 2.5 days for the Reading task.

Training Process. Since we have a relatively small dataset, we apply a cross-validation strategy and report averaged performance. All the models are trained using 10-fold cross-validation over 10 runs. More specifically, we conduct the cross-validation across users to ensure there is no user contributing data points to both the training and test sets of a given fold. In addition, cross-validation is also stratified so that the distribution of data points in each fold is kept similar to that of the dataset.

Evaluation Metrics. We report the same evaluation metrics used in previous work (Sriram et al., 2023; Jang et al., 2021), which includes: (i) AUC (Area Under Curve): it measures the accuracy of the classifier in distinguishing between patients and

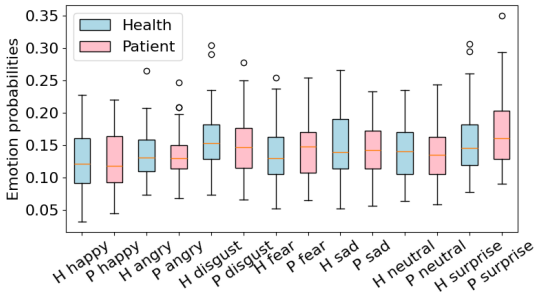

Figure 2: AD patients (pink) and healthy controls (blue) emotion distribution on Picture Description task.

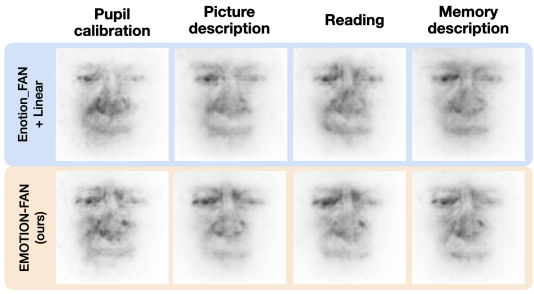

Figure 3: Visualization results from two models: EMOTION-FAN+Linear and EMOTION-FAN (ours) on four tasks.

controls. (ii) Sensitivity (or true positive rate): this indicates the model's ability to detect patients. (iii) Specificity (or true negative rate): it shows the model's ability to detect controls. The AUC provides an overall performance measure. Sensitivity and specificity are important in medical applications to estimate the likelihood of false negatives and false positives.

4.3. Results and Analysis

In Table 1, we show the performance of our two emotion-based models, EMOTION-FAN+linear and EMOTION-FAN (ours), on four tasks: Pupil Calibration, Picture Description, Reading, and Memory. Additionally, since facial asymmetry increases with age (Chien et al., 2023), we verify whether age can by itself be a strong predictor of AD classification. For this, we train a one-layer classifier using ReLU activation function, the results are shown in the first row of Table 1.

When classification is done directly from the emotion outputs, all three models perform poorly, with AUC of 56–68%, Sensitivity of 66–70%, and Specificity of 65–73%. We suspect that this is due to the tasks themselves not inherently eliciting strong emotions in our participants. Figure 2 supports this hypothesis, showing an even distribution of predicted emotions for both groups in the Picture Description task. This suggests the minimal difference in emotional response between AD patients and healthy controls. Similar trends were also observed in the other two tasks (see Appendix C).

On the other hand, our fine-tuned EMOTION-FAN model outperforms the EMOTION-FAN+linear model, achieving 20–25% improvement in AUC score, 15–18% improvement in sensitivity, and 12–15% improvement in specificity across all tasks. To understand how our fine-tuned model outperforms the original EMOTION-FAN, we visualize the integrated gradients[1] of different models for each task, as shown in Figure 3. For each task, we first average across the frames in each video and then average all data points in each task. As a result, we are able to show the saliency parts of two models in different tasks. Plausibly, this is an indication that the model is capturing facial asymmetry which has been connected to the presence of dementia (Chien et al., 2023).

5. AD Classification Using Facial and Eye-tracking (ET) Data

Building on the promising results from facial patterns in Section 4, this section investigates the potential of combining facial and eye-tracking (ET) data for more accurate AD classification. We obtain classification results from eye movement by replicating the experiments of Sriram et al. (2023). Their approach uses a classifier, called VTNet, which combines a CNN for handling ET scanpaths and a RNN for processing ET sequences. For facial data, we employ the fine-tuned EMOTION-FAN model described in Section 4.

This section is organized as follows: in Section 5.1, we present ET data processing as well as the VT-Net model. We then implement a simple late fusion method in Section 5.2 that has been shown to be effective in previous work (Barral et al., 2020; Jang

1. We employ the integrated gradients visualization from Captum package: https://captum.ai/tutorials/TorchVision_Interpret.

Table 1: AD classification results using facial data on four tasks. Age (baseline): AD classification using participants' age. We show results (averaged over 10-run 10-fold cross-validation) using EMOTION-FAN+linear and EMOTION-FAN (ours) on each task. The best scores are bolded.

Task	Features	Models	AUC	Sensitivity	Specificity
-	*Age (baseline)*	1-layer neural network	0.55 ± 0.05	0.64 ± 0.06	0.63 ± 0.05
Pupil Calibration	Emotions	EMOTION-FAN+linear	0.56 ± 0.03	0.66 ± 0.04	0.65 ± 0.05
	Facial patterns	EMOTION-FAN (ours)	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.84 \pm 0.03}$	$\mathbf{0.80 \pm 0.04}$
Picture Description	Emotions	EMOTION-FAN+linear	0.57 ± 0.04	0.64 ± 0.06	0.67 ± 0.06
	Facial patterns	EMOTION-FAN (ours)	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.78 \pm 0.03}$
Reading	Emotions	EMOTION-FAN+linear	0.68 ± 0.06	0.70 ± 0.06	0.73 ± 0.05
	Facial patterns	EMOTION-FAN (ours)	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.81 \pm 0.02}$
Memory	Emotions	EMOTION-FAN+linear	0.61 ± 0.03	0.67 ± 0.05	0.68 ± 0.05
	Facial patterns	EMOTION-FAN (ours)	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.77 \pm 0.03}$	$\mathbf{0.82 \pm 0.03}$

et al., 2021). Lastly, in Section 5.3, we compare the results from using single and multiple modalities, as well as showing an in-depth error analysis.

5.1. AD Classification Using ET Data

ET Data Representation. We collect individual ET samples at the sampling rate of 120Hz. Each sample comprises a six-dimensional vector, including the gaze coordinates, the distances of both the left and right eyes from the screen (which helps estimate the distance of the head from the screen), and the pupil sizes of both the left and right eyes (details in Appendix D).

VTNet Model. The *VTNet* model, first introduced in Sims and Conati (2020), was designed to identify user confusion using raw ET data. It comprises a single-layer GRU and a two-layer CNN. The GRU processes raw ET sequences while the CNN handles the spatial representation, or scanpath, indicating fixation locations and transitions. A self-attention layer is later added before the GRU in Sriram et al. (2023) to focus on important parts of the sequential ET sequences, thereby effectively capturing long-term dependencies. The dimension of this self-attention layer is set to 6 to match the dimensionality of the gaze data, and the number of parallel attention heads is set to 1 to maintain model simplicity and computational efficiency, especially important given the limit size of the dataset (Jang et al., 2021). The GRU's output, a 256-unit hidden state, is concatenated with a 50-element vector from the CNN to form a 306-sized vector. This vector is then sent as input to a simple neural network featuring one hidden layer and a SoftMax layer that outputs two confidence levels, classifying the input as either AD or control. We use the top performing VTNet model in (Sriram et al., 2023) for the following comparisons and combinations.

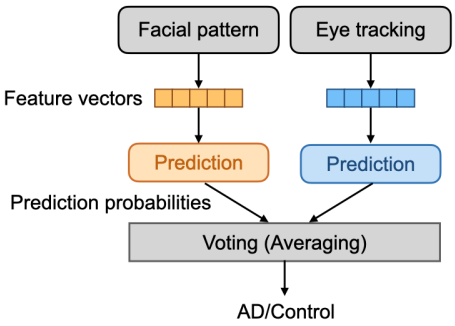

Figure 4: Fusion Model Overview.

5.2. Fusion Method

To combine different modalities for AD classification, we explore a late fusion method which is generally employed for multimodal data (Jang et al., 2021; Zheng et al., 2023; Barral et al., 2020). The late fusion method aggregates predictions from ET and video modalities at the decision level, as shown in Figure 4. Following the work in Barral et al. (2020), we adopt a late fusion approach known as "voting by averaging". The predictions for each data point are initially obtained from the best performing models in different modalities, i.e., fine-tuned EMOTION-FAN model for video and VTNet model for eye-tracking. A voting (averaging) mechanism is then used to determine the

Table 2: Pupil Calibration, Picture Description, and Reading results using ET, video, and ET+Video modalities. Combined†: the averaged results over all tasks with ET+Video modality.

Task	Modality	Models	AUC	Sensitivity	Specificity
Pupil Calibration	ET	VTNet (Sriram et al., 2023)	0.78 ± 0.01	0.71 ± 0.02	0.75 ± 0.01
	Video	EMOTION-FAN (ours)	0.81 ± 0.02	0.84 ± 0.03	0.80 ± 0.04
	ET+Video	VTNet + EMOTION-FAN (ours)	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.81 \pm 0.03}$
Picture Description	ET	VTNet (Sriram et al., 2023)	0.76 ± 0.01	0.70 ± 0.02	0.73 ± 0.02
	Video	EMOTION-FAN (ours)	0.79 ± 0.02	0.82 ± 0.03	0.78 ± 0.03
	ET+Video	VTNet + EMOTION-FAN (ours)	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.82 \pm 0.03}$
Reading	ET	VTNet (Sriram et al., 2023)	0.78 ± 0.01	0.70 ± 0.01	0.80 ± 0.02
	Video	EMOTION-FAN (ours)	0.83 ± 0.02	0.83 ± 0.03	0.81 ± 0.02
	ET+Video	VTNet + EMOTION-FAN (ours)	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.86 \pm 0.03}$
Combined†	ET+Video	VTNet + EMOTION-FAN (ours)	0.88 ± 0.01	0.85 ± 0.03	0.86 ± 0.03

Table 3: MANOVA results on three tasks. All the scores are significant (i.e., $p < 0.001$)

Task	**Overall**			**AUC**		**Sensitivity**		**Specificity**	
	$F_{6,50}$	partial η^2	Wilk's Λ	$F_{2,27}$	partial η^2	$F_{2,27}$	partial η^2	$F_{2,27}$	partial η^2
Pupil Calibration	29.19	0.78	0.048	31.66	0.70	58.13	0.81	28.42	0.68
Picture Description	29.61	0.78	0.049	40.37	0.75	55.65	0.81	11.34	0.46
Reading	42.84	0.84	0.027	126.55	0.90	135.63	0.91	20.89	0.61

final prediction, integrating inputs from both modalities. Despite its simplicity, this late fusion method has shown its effectiveness in the literature (Battiti and Colla, 1994; Jang et al., 2021). In addition, changes in decision-making after the fusion are easy to interpret and track, as we will discuss in the next section.

5.3. Results and Analysis

Main Findings In Table 2, we show the best results for single modality (eye-tracking, video) and the fused ET+video modal on all three tasks, with best performance in bold. The last row (combined) shows the micro-averaged fusion results for all tasks and modalities. Note that we did not include the fused results for the Memory Description task since this task had no visual stimulus during the experiments. Therefore, no eye-tracking data could be used for the fusion method. Looking closely at Table 2, we see that video data consistently outperforms eye-tracking across all three tasks. Specifically, the AUC score improves by 3–5%, the sensitivity score by 12–13%, and the specificity score by 1–5%. We observe a significant improvement in sensitivity, which suggests that our facial model performs well in recognizing patients with AD. Notably, the fusion model achieves even higher performances exceeding the video-only modality by 3–5% in AUC, 1–3% in sensitivity, and 1–5% in specificity.

To formally compare these results, we conduct statistical analysis on each task. We first use a one-way MANOVA test, where the classifier type (VTNet, EMOTION-FAN (ours), and VTNet+EMOTION-FAN (ours)) is the independent variable and the three performance metrics (AUC, sensitivity, and specificity scores) serve as dependent variables. As shown in Table 3, for all three tasks, the MANOVA analysis reveals a significant effect ($p < 0.001$) of the classifier type on all three performance metrics. For post-hoc comparisons, we use the Tukey's HSD test. Pair-wise comparison results indicate that across all tasks, the fusion model (VTNet+EMOTION-FAN (ours)) achieves statistically higher AUC scores than the video-only model (EMOTION-FAN (ours)), which in turn significantly outperforms the eye-tracking-only model (VTNet). This ranking also holds for sensitivity in the Reading task and for specificity in the Picture Description task. However, for the sensitivity scores in the Pupil Calibration and Picture Description tasks, as well as the specificity scores in the Pupil Calibration task, the fusion model is found to be equivalent to the video-only model, and both models statistically outperform the eye-tracking-only model. Finally, in terms of specificity for the Reading task, no significant

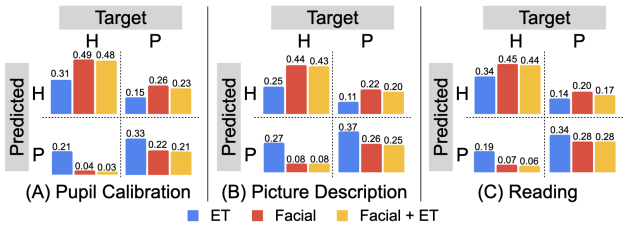

Figure 5: Comparison of confusion matrices using different models. Blue: VTNet with ET data; Red: EMOTION-FAN (ours) with facial data; Yellow: model with Facial+ET data. H: Healthy. P: Patient.

difference is observed between the video-only model and the eye-tracking-only model; however, the fusion model outperforms both.

These results strongly suggest that learning from multimodal data sources improves the model's performance across all tasks. Additionally, we calculate the combined performance by averaging the results over all three tasks. The outcomes are comparable to those of the Reading task, indicating that the information across tasks might not vary significantly.

Error Analysis To better understand how the fusion model improves the performance compared with the facial or ET model alone, we conduct an analysis before and after the fusion, as shown in Figure 5, with detailed scores reported in Appendix E. In this Figure, we present the confusion matrices: true positive (patient-patient, or P-P), true negative (health-health, or H-H), false positive (health-patient, H-P), and false negative (patient-health, P-H) on three settings (Facial, ET, Facial+ET) and on all three tasks. The scores are averaged over 10-run 10-cross-validation.

In the Pupil Calibration task, our facial model greatly outperforms the ET model in identifying healthy controls (or true negatives, 49% vs. 31%). The false negative rate is also much lower (4% vs. 21%). When the results from the facial model and the ET model are combined, the good performance of the true negative and false negative rates is comparable to the results from the facial model alone (48% vs. 49%, 3% vs. 4%). This implies that the facial model has strong predictive power and can lead to good aggregate results. In comparison, the ET model achieves better results than the facial model in categorizing patients (or true positives, 33% vs. 22%) while also having a lower false positive rate (15% vs. 26%). Combing these results help correct for the false positive rate by 3% while the true positive rate does not show improvement. In both the Picture Description task and the Reading task, we observed similar trends to the Pupil Calibration task for all four metrics for all three models. Interestingly, in the Reading task, our facial model outperforms the Picture Description task in recognizing patients with AD, which in turn outperforms the Pupil Calibration task. Since our facial model is based on a pre-trained emotion recognition model, the EMOTION-FAN model performs better for tasks that elicit more emotions.

From the above analysis, we can see that the models trained on facial and ET data show complementary effects, leading to better fusion results. Nevertheless, the effect of improving the true positive rate is not significant ($\pm 1\%$ for all tasks). Thus, there is a great potential to improve the true positive rate using other fusion strategies, which we will explore in future work.

6. Conclusion and Future Work

In this paper, we investigate the use of video modality in AD classification. We extract facial patterns related to emotions by leveraging the EMOTION-FAN model, either by directly using it or fine-tuning it for our task. Experimental results show that the fine-tuned emotion model greatly outperforms the model using eye-tracking data, validating our hypothesis that facial patterns provide valuable information that can improve accuracy. Moreover, we conduct a late fusion strategy to combine the predictions from ET and video data. The fused result across all three tasks underscores the benefits of multimodal data integration.

For future work, we plan to leverage more powerful vision encoders. Currently, the EMOTION-FAN model employs ResNet18 to encode video into frame features. However, emerging vision encoders like Vision Transformer (Dosovitskiy et al., 2020) offer enhanced capabilities and could potentially improve visual feature extraction. We also intend to explore alternative fusion approaches, such as combining the feature embeddings prior to the decision-making or integrating features from two modalities right after processing by the CNNs. This could provide more effective ways to leverage multimodal data. Additionally, it is intriguing to explore visualization techniques such as LIME (Ribeiro et al., 2016) to analyze which specific parts of the faces in a video frame the model focuses on in order to improve the interpretability of the model. Lastly, it will be worthwhile to compare with speech models and potentially integrate the speech modality in AD classification.

References

Kenneth Asplund, Astrid Norberg, Rolf Adolfsson, and Howard M Waxman. 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.

Oswald Barral, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. Non-invasive classification of alzheimer's disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR, 2020.

Roberto Battiti and Anna Maria Colla. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 1994.

Juan Biondi, Gerardo Fernandez, Silvia Castro, and Osvaldo Agamennoni. Eye-movement behavior identification for ad diagnosis. *arXiv preprint arXiv:1702.00837*, 2017.

Weiwei Cai, Ming Gao, Runmin Liu, and Jie Mao. Mifad-net: Multi-layer interactive feature fusion network with angular distance loss for face emotion recognition. *Frontiers in psychology*, pages 762795–762795, 2021.

Ching-Fang Chien, Jia-Li Sung, Chung-Pang Wang, Chen-Wen Yen, and Yuan-Han Yang. Analyzing facial asymmetry in alzheimer's dementia using image-based technology. *Biomedicines*, 11(10):2802, 2023.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 653–656, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Kathleen C Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. Predicting mci status from multimodal language data using cascaded classifiers. *Frontiers in aging neuroscience*, 11:205, 2019.

Siobhan Garbutt, Alisa Matlin, Joanna Hellmuth, Ana K Schenk, Julene K Johnson, Howard Rosen, David Dean, Joel Kramer, John Neuhaus, Bruce L Miller, et al. Oculomotor function in frontotemporal lobar degeneration, related disorders and alzheimer's disease. *Brain*, 131(5):1268–1281, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016b. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Hyeju Jang, Thomas Soroski, Matteo Rizzo, Oswald Barral, Anuj Harisinghani, Sally Newton-Mason, Saffrin Granby, Thiago Monnerat Stutz da Cunha Vasco, Caitlin Lewis, Pavan Tutt, et al. Classification of alzheimer's disease leveraging multi-task machine learning analysis of speech and eye-

movement data. *Frontiers in Human Neuroscience*, 2021.

Patrik Jonell, Birger Moëll, Krister Håkansson, Gustav Eje Henter, Taras Kucherenko, Olga Mikheeva, Göran Hagman, Jasper Holleman, Miia Kivipelto, Hedvig Kjellström, et al. Multimodal capture of patient behaviour for improved detection of early dementia: clinical feasibility and preliminary results. *Frontiers in Computer Science*, 3:642633, 2021.

Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

Weirui Kong, Hyeju Jang, Giuseppe Carenini, and Thalia Field. A neural model for predicting dementia from language. In *Machine Learning for Healthcare Conference*, pages 270–286. PMLR, 2019.

Michael R MacAskill and Tim J Anderson. Eye movements in neurodegenerative diseases. *Current opinion in neurology*, 29(1):61–68, 2016.

Masateru Matsushita, Yusuke Yatabe, Asuka Koyama, Akiko Katsuya, Daisuke Ijichi, Yusuke Miyagawa, Hiroto Ikezaki, Noboru Furukawa, Manabu Ikeda, and Mamoru Hashimoto. Are saving appearance responses typical communication patterns in alzheimer's disease? *Plos one*, 13(5):e0197468, 2018.

Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

Robert J Molitor, Philip C Ko, and Brandon A Ally. Eye movements in alzheimer's disease. *Journal of Alzheimer's disease*, 44(1):1–12, 2015.

Ivanna M Pavisic, Nicholas C Firth, Samuel Parsons, David Martinez Rego, Timothy J Shakespeare, Keir XX Yong, Catherine F Slattery, Ross W Paterson, Alexander JM Foulkes, Kirsty Macpherson, et al. Eyetracking metrics in young onset alzheimer's disease: a window into cognitive visual functions. *Frontiers in neurology*, 8:377, 2017.

María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer's disease and automatic speech analysis: a review. *Expert systems with applications*, 150:113213, 2020.

Eric M Reiman, Jessica B Langbaum, Pierre N Tariot, Francisco Lopera, Randall J Bateman, John C Morris, Reisa A Sperling, Paul S Aisen, Allen D Roses, Kathleen A Welsh-Bohmer, et al. Cap—advancing the evaluation of preclinical alzheimer disease treatments. *Nature Reviews Neurology*, 12(1):56–61, 2016.

Maria Revi. Alzheimer's disease therapeutic approaches. *GeNeDis 2018: Genetics and Neurodegeneration*, pages 105–116, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Douglas W Scharre. Preclinical, prodromal, and dementia stages of alzheimer's disease. *Pract Neurol*, 15:36–47, 2019.

Ulrich Seidl, Ulrike Lueken, Philipp A Thomann, Andreas Kruse, and Johannes Schröder. Facial expression in alzheimer's disease: Impact of cognitive deficits and neuropsychiatric symptoms. *American Journal of Alzheimer's Disease & Other Dementias®*, 27(2):100–106, 2012.

Zhengyan Sheng, Zhiqiang Guo, Xin Li, Yunxia Li, and Zhenhua Ling. Dementia detection by fusing speech and eye-tracking representation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6457–6461. IEEE, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Shane D Sims and Cristina Conati. A neural architecture for detecting user confusion in eye-tracking data. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 15–23, 2020.

MC Smith. Facial expression in mild dementia of the alzheimer type. *Behavioural Neurology*, 8:149–156, 1995.

Harshinee Sriram, Cristina Conati, and Thalia Field. Classification of alzheimer's disease with deep learning on eye-tracking data. In *Proceedings of the 25th*

International Conference on Multimodal Interaction, pages 104–113, 2023.

Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006.

Yumi Umeda-Kameyama, Masashi Kameyama, Tomoki Tanaka, Bo-Kyung Son, Taro Kojima, Makoto Fukasawa, Tomomichi Iizuka, Sumito Ogawa, Katsuya Iijima, and Masahiro Akishita. Screening of alzheimer's disease by facial complexion using artificial intelligence. *Aging (Albany NY)*, 13(2):1765, 2021.

WHO. Dementia — who.int. https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed 09-05-2024].

Chuheng Zheng, Mondher Bouazizi, Tomoaki Ohtsuki, Momoko Kitazawa, Toshiro Horigome, and Taishiro Kishimoto. Detecting dementia from face-related features with automated computational methods. *Bioengineering*, 10(7):862, 2023.

Appendix A. Dataset Details

Table 4 shows the participant demographics in the dataset and Figure 6 provides the overview of four tasks originated from Jang et al. (2021).

Table 4: Participant Demographics. Parti. = Participants, M = Male, F = Female, MoCA = Montreal Cognitive Assessment Score.

Group	Parti.	Age	Gender	MoCA
Control	75	62 ± 15	22M/53F	27 ± 3
Patients	69	72 ± 9	33M/36F	18 ± 7

Appendix B. Data Preprocess

We show the data preprocessing in Figure 7.

Appendix C. Emotion Distribution

Figure 8, Figure 9, and Figure 10 shows the emotion distribution on Pupil Calibration, Picture Description and Reading tasks, respectively. We observe evenly distributed emotions, suggesting that the differences in emotional responses are not predictive for AD classification.

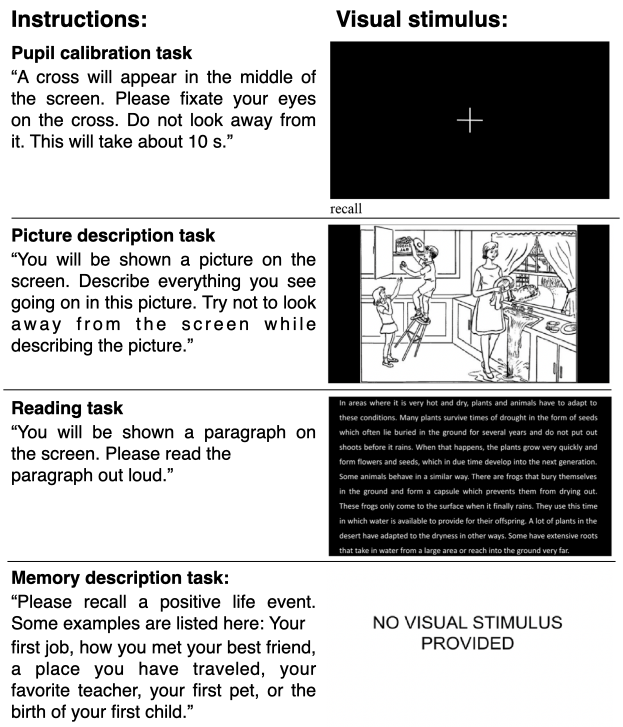

Figure 6: An overview for the four tasks originated from Jang et al. (2021) and employed in our paper. Left: instructions given for the participants. Right: visual stimulus.

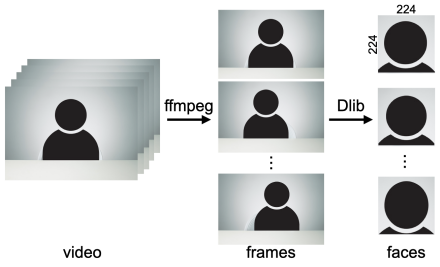

Figure 7: Data preprocessing. We use FFmpeg (Tomar, 2006) to cut the videos into images. The image frames are then processed with the Dlib toolbox for face detection.

Appendix D. ET Data Representation

Figure 11 shows how the raw ET samples from each user are represented by a 2D array. In this array, rows correspond to individual ET samples collected at the sampling rate of 120Hz. Each sample comprises a six-dimensional vector, including the gaze coordinates

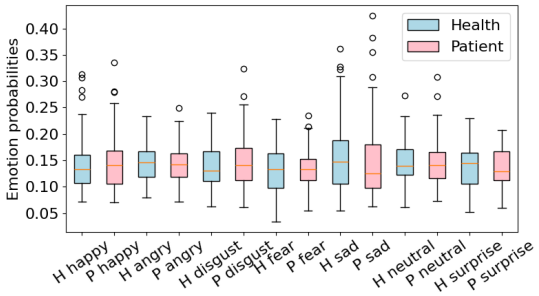

Figure 8: AD patients (pink) and healthy controls (blue) emotion distribution on Pupil Calibration task.

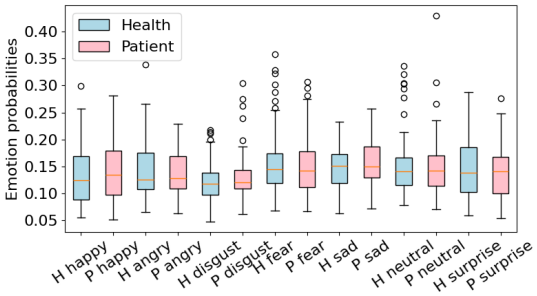

Figure 10: AD patients (pink) and healthy controls (blue) emotion distribution on Reading task.

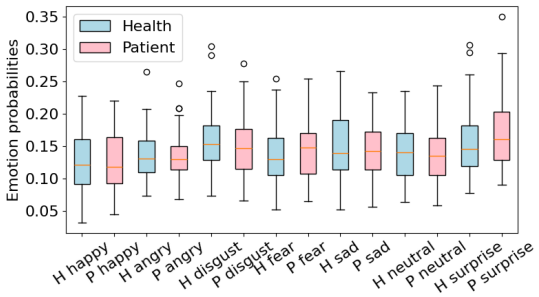

Figure 9: AD patients (pink) and healthy controls (blue) emotion distribution on Picture Description task.

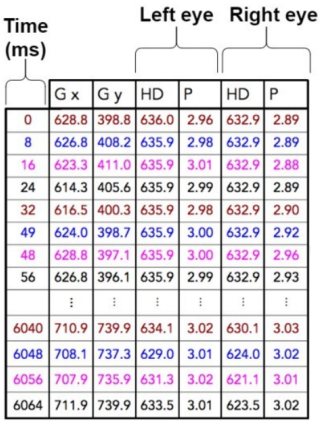

Figure 11: An example of an ET datapoint, which is a sequence of ET samples (rows).

(G_x, G_y), the distances (HD) of both the left and right eyes from the screen (which helps estimate the distance of the head from the screen), and the pupil sizes of both the left and right eyes (P).

Appendix E. Error Analysis

We conducted the error analysis of the fusion method, with the results shown in Table 5. The scores are highlighted in different colors to represent four metrics: true positive, false positive, false negative, and true negative. Additionally, the ↑ arrows indicate that higher scores are preferable, while the ↓ arrows denote that lower scores are preferable. 'H' represents healthy controls, and 'P' denotes patients. Scores are presented as mean ± standard deviation. The table shows that the fusion model significantly improves the true and false positive rates, particularly in comparison to using the ET model alone. Similarly, the false negative and true negative rates benefit from the strengths of the ET model, enhancing overall performance after the fusion. These phenomena indicate the effectiveness of the fusion method when predicting the AD/control.

Table 5: Model prediction using Facial, ET, and Facial+ET modalities on three tasks. Target: gold class. H: Healthy. P: Patient. We color four metrics: true positive, true negative, false positive, false negative. The ↑ and ↓ indicate resp. expected higher score and lower score in each metric. Scores are averaged (mean ± standard deviation).

	Pupil Calibration					
Target	ET		Facial		Facial+ET	
H	H ↑	0.31±0.01	H ↑	0.49±0.01	H ↑ P	0.48±0.01 0.01±0.00
H	P ↓	0.21±0.01	P ↓	0.04±0.01	H P ↓	0.01±0.01 0.03±0.01
P	H ↓	0.15±0.01	H ↓	0.26±0.03	H ↓ P	0.23±0.02 0.03±0.01
P	P ↑	0.33±0.01	P ↑	0.22±0.03	H P ↑	0.01±0.01 0.21±0.03
	Picture Description					
Target	ET		Facial		Facial+ET	
H	H ↑	0.25±0.02	H ↑	0.44±0.02	H ↑ P	0.43±0.02 0.00±0.00
H	P ↓	0.27±0.02	P ↓	0.08±0.02	H P ↓	0.00±0.00 0.08±0.02
P	H ↓	0.11±0.03	H ↓	0.22±0.01	H ↓ P	0.20±0.01 0.02±0.01
P	P ↑	0.37±0.03	P ↑	0.26±0.01	H P ↑	0.01±0.00 0.25±0.02
	Reading					
Target	ET		Facial		Facial+ET	
H	H ↑	0.34±0.04	H ↑	0.45±0.01	H ↑ P	0.44±0.01 0.01±0.01
H	P ↓	0.19±0.04	P ↓	0.07±0.01	H P ↓	0.01±0.01 0.06±0.01
P	H ↓	0.14±0.02	H ↓	0.20±0.02	H ↓ P	0.17±0.01 0.03±0.02
P	P ↑	0.34±0.02	P ↑	0.28±0.02	H P ↑	0.00±0.01 0.28±0.02