Automatic Detection of Cognitive Impairment Through Facial Emotion Analysis

applied sciences

MDPI

Article

Automatic Detection of Cognitive Impairment Through Facial Emotion Analysis

Letizia Bergamasco [1,2], Federica Lorenzo [1], Anita Coletta [1], Gabriella Olmo [1,*], Aurora Cermelli [3,4], Elisa Rubino [3,4] and Innocenzo Rainero [3,4]

[1] Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy; letizia.bergamasco@polito.it (L.B.); federica.lorenzo@studenti.polito.it (F.L.); anita.coletta@polito.it (A.C.)
[2] LINKS Foundation, 10138 Turin, Italy
[3] Center for Alzheimer's Disease and Related Dementias, Department of Neuroscience and Mental Health, A.O.U. Città della Salute e della Scienza di Torino, 10126 Turin, Italy; aurora.cermelli@unito.it (A.C.); elisa.rubino@unito.it (E.R.); innocenzo.rainero@unito.it (I.R.)
[4] Department of Neuroscience "Rita Levi Montalcini", University of Torino, 10126 Turin, Italy
* Correspondence: gabriella.olmo@polito.it

Abstract

Altered facial expressivity is frequently recognized in cognitively impaired individuals. This makes facial emotion identification a promising tool with which to support the diagnostic process. We propose a novel, non-invasive approach for detecting cognitive impairment based on facial emotion analysis. We design a protocol for emotion elicitation using visual and auditory standardized stimuli. We collect facial emotion video recordings from 32 cognitively impaired and 28 healthy control subjects. To track the evolution of emotions during the experiment, we train a deep convolutional neural network on the AffectNet dataset for emotion recognition from facial images. Emotions are described using a dimensional affect model, namely the continuous dimensions of valence and arousal, rather than discrete categories, enabling a more nuanced analysis. The collected facial emotion data are used to train a classifier to distinguish cognitively impaired and healthy subjects. Our k-nearest neighbors model achieves a cross-validation accuracy of 76.7%, demonstrating the feasibility of automatic cognitive impairment detection from facial expressions. These results highlight the potential of facial expressions as early markers of cognitive impairment, which could enhance non-invasive screening methods for early diagnosis.

Keywords: artificial intelligence; cognitive impairment detection; dementia; facial emotion recognition; mild cognitive impairment

Academic Editor: Jing Jin

Received: 21 July 2025
Revised: 13 August 2025
Accepted: 14 August 2025
Published: 19 August 2025

Citation: Bergamasco, L.; Lorenzo, F.; Coletta, A.; Olmo, G.; Cermelli, A.; Rubino, E.; Rainero, I. Automatic Detection of Cognitive Impairment Through Facial Emotion Analysis. *Appl. Sci.* **2025**, *15*, 9103. https://doi.org/10.3390/app15169103

1. Introduction

Cognitive impairment (CI) refers to a decline in cognitive functions, such as memory, attention, language, or problem-solving. It encompasses a continuum of conditions, ranging from mild cognitive impairment (MCI), where cognitive changes are clinically appreciable but do not significantly affect daily activities, to forms of overt dementia, involving substantial cognitive decline and interfering with the person's independence. With more than 10 million new cases each year worldwide, dementia is one of the most impactful syndrome in modern society at a global level [1]. The most common form is Alzheimer's disease (AD); however, other forms such as vascular, frontotemporal, dementia with Lewy bodies, or mixed forms have a relevant social and clinical impact, and their differential diagnosis is often difficult.

Current therapies focus on providing temporary symptom relief but have little to no effectiveness in modifying disease progression [2]. However, the Food and Drug Administration has recently approved disease-modifying therapies for AD [3,4], which are effective only if administered in the very initial, possibly preclinical, phases of the disease. Hence, early detection of dementia is crucial for including subjects in clinical trials and improving the quality of life of patients and their caregivers through proper lifestyle modifications.

The diagnosis of CI relies on a combination of medical history, neuropsychological assessment, neuroimaging, and lab tests, including a lumbar puncture to look for AD biomarkers in the cerebrospinal fluid–beta-amyloid ($A\beta 42$), $A\beta 42/A\beta 40$ ratio, total tau, and phosphorylated tau proteins [5]. These techniques are often expensive, possibly invasive, and require specialized healthcare professionals. Hence, the development of cost-effective techniques for early CI detection is potentially very important.

It is well known that facial expressions encompass relevant information related to the cognitive status of the individual. They are controlled by complex cerebral circuits and convey various types of messages, above all those related to the emotional state. Altered facial expressivity is frequently recognized in cognitively impaired individuals. Alterations tend to be related to forms and stages of dementia [6,7]; this makes facial emotion identification a promising tool also for differential dementia diagnosis.

Direct evaluation of facial expressions is complex and operator-dependent. In cognitively impaired patients, such evaluation can be hindered by a lack of cooperation, and small yet significant details may go unnoticed by the examiner. In addition, the emotional dimension is currently underexplored in common neuropsychological tests such as the Mini Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA). On the other hand, artificial intelligence (AI) approaches, especially deep learning (DL)-based techniques, have great potential in the field of facial expression analysis and may be applied to CI detection, thus contributing to an earlier and more accurate etiological diagnostic process. Encouraging outcomes in this regard have already been reported in [8–10].

This work aims to develop an AI-based system for detecting cognitive impairment through facial emotion analysis, moving in the direction of identifying non-invasive CI screening methods. The main contributions are summarized as follows:

- We address the important problem of detecting cognitive impairment from its very beginning (hence, including MCI) using non-invasive and cost-effective techniques.
- We validate the assumption that facial expressions in response to emotional elicitation are different in cognitively impaired and healthy subjects.
- To this end, we propose an AI framework based on facial emotion data, using a dimensional model of affect. We design and test an emotion elicitation protocol based on standardized stimuli.
- We test our system on video recordings collected from cognitively impaired and healthy control subjects, with a ground truth classification of CI based on a comprehensive neurological and neurocognitive assessment.

In this paper, Section 2 contains a brief review of already published papers on CI detection using facial expressions and emotion recognition. Section 3 details the employed data collection protocol, and the CI detection implementation; Sections 4 and 5 present and discuss the obtained results; lastly, Section 6 concludes the paper and outlines possible directions for future investigations.

2. Background

2.1. Deep Learning for CI Detection Using Facial Features

A recent review paper by Alsuhaibani et al. [11] explored emerging DL approaches for non-invasive CI detection. The authors analyzed the use of speech, facial and motor indicators, concluding that, while speech-based methods provide high performance, facial expression analysis is promising but needs further investigation to ensure proper robustness.

Some studies attempted to detect cognitive decline from raw images. Sun et al. [8] worked on the I-CONECT dataset [12], containing semi-structured interviews of 186 participants. They selected four conversational themes involving 147 subjects (83 MCI and 64 healthy controls - HC) and used facial videos to train a multi-branch classifier–video vision transformer (MC-ViViT) model, achieving 90.63% accuracy in distinguishing MCI from HC for one of the selected conversational themes. Umeda-Kameyama et al. [9] claimed to achieve a 92.56% accuracy in CI detection using an Xception DL model trained on a dataset of 484 face images (121 dementia patients, 117 HC). However, the authors recognized that the results might be affected by institutional biases.

Other studies have focused on extracting proper features from face images. Zheng et al. [13] used face mesh, histograms of oriented gradients (HOGs), and action units (AUs—see Section 2.2 for formal definition) in an attempt to mitigate the bias caused by varying lighting conditions or data collection environments. They reached 79% accuracy on dementia detection, using a long short-term memory (LSTM) model trained on HOG features extracted from a section of video data from the PROMPT dataset [14]. This dataset encompasses 447 videos of 117 subjects, including HC and various pathological individuals affected by dementia, bipolar disorder, and depression. However, Zheng [13] also reported a potential institutional bias, since HOG features are sensitive to light changes, and healthy and dementia data were collected in different environmental conditions. With AUs and face mesh features, the classification performance was 71% and 66%, respectively.

Few studies have focused on facial emotions for the development of automatic CI detection. To show that cognitively impaired subjects express facial emotions differently from cognitively unimpaired ones, Jiang et al. [15] conducted a study involving 493 participants encompassing HC and individuals with CI of varying severity and etiology. They analyzed the facial emotions of participants during a memory test using a DL model for facial emotion recognition and provided evidence that cognitively impaired subjects display less positive emotions, more negative emotions, and increased facial expressiveness. Fei et al. [10] presented a system to detect CI through the analysis of categorical facial emotions. The system included three main components: an interface to display video emotional stimuli and record facial expressions; a DL-based model to extract an emotion evolution matrix from video frames; and a support vector machine (SVM) classifier to distinguish cognitively impaired and HC subjects. While the purpose of this work is similar to ours, Fei [10] adopts a categorical approach for emotion representation. On the other hand, the integration of a dimensional model of affect and standardized emotion elicitation stimuli represents a significant improvement, as discussed in the following section.

2.2. Automated Facial Emotion Recognition

Three main models are used to objectively represent human emotions. The first and most employed is the categorical model, where emotions are represented by a list of discrete categories. For example, Ekman's basic emotions model [16] considers six basic emotions—anger, disgust, fear, happiness, sadness and surprise—plus the neutral state. Alternatively, Plutchik's model includes eight primary emotions grouped into polar opposites: joy and sadness, acceptance and disgust, fear and anger, and surprise and anticipation. Emotions are displayed in a flower-shaped representation with eight petals, known as Plutchik's

Wheel of Emotions [17], and they intensify moving from the outside to the center. However, even taking into account the intensity of the primary emotions, categorical models cannot fully represent the nuance and complexity of affective behaviors.

The facial action coding system (FACS) [18] is a system in which facial expressions are decomposed into single muscle movements (AUs). Every facial expression can be coded using a combination of AUs. However, due to several factors such as lighting changes, position variations, and differences among individuals, AU recognition is difficult. Moreover, AU annotation is expensive and time-consuming, and this limits the availability of AU datasets.

Since categorical models cannot adequately describe mixed emotions, researchers have proposed to represent affective behaviors through continuous dimensions. Among all dimensional models, the circumplex model [19] is widely used. Emotions are expressed as points in a two-dimensional space, whose perpendicular axes represent valence (positive or negative emotional state) and arousal (the strength of emotion activation). Overall, dimensional models are more powerful than categorical ones in capturing all possible emotion nuances. Nevertheless, few works employ dimensional models for automated emotion recognition. This may be due to the high cost of building a large database and covering the continuous space of valence and arousal; in fact, there is a scarcity of annotated face databases in the continuous domain [20].

Commonly used DL models in facial emotion recognition (FER) are convolutional neural networks (CNNs), because of their ability to automatically learn relevant features directly from facial images, without the need for explicit feature engineering. CNN models have already been used to classify emotions according to categorical [21,22] or seldom dimensional models [20,23,24]. Moreover, numerous recent studies have made use of the attention mechanism, which proved to be effective in FER [25,26]. Most papers have applied the attention mechanism to CNNs in discrete emotion classification, but Xiaohua et al. [27] have also demonstrated successful application in predicting valence and arousal using a bi-directional recurrent neural network with self-attention.

As highlighted in a recent survey by Karnati [28], the development of robust FER systems still faces several challenges, including pose variation, occlusions, illumination changes, noisy labels, and overfitting due to limited and imbalanced datasets. These limitations are particularly relevant in real-world and clinical applications and should be carefully considered when designing or deploying emotion recognition models.

3. Materials and Methods

3.1. System Overview

This work proposes an automated system to distinguish cognitively impaired patients from healthy individuals, based only on facial emotions. It consists of four main parts. (i) A dimensional CNN model for FER is trained on the AffectNet dataset [20]. (ii) A set of individuals, encompassing cognitively impaired subjects and HC, undergo a properly designed emotion elicitation protocol while their facial expressions are video-recorded. (iii) The trained CNN model is applied to the collected video dataset to obtain the temporal evolution of emotions. (iv) The facial emotion data are used to train a machine learning (ML) model devoted to CI detection. An overview of the pipeline is depicted in Figure 1.

The CNN training is implemented on a local server equipped with NVIDIA GeForce RTX 3080 GPU, Intel i9-10900X CPU and 64 GB RAM. The utilized versions of Python, and Tensorflow and Keras [29] are 3.10.16 and 2.9.0, respectively, with CUDA 12.2. The design and implementation of the emotion elicitation video makes use of PsychoPy v2022.2.4, an open-source package for running behavioral sciences experiments in Python [30]. The ML classifier of CI versus HC is realized using the scikit-learn Python library [31].

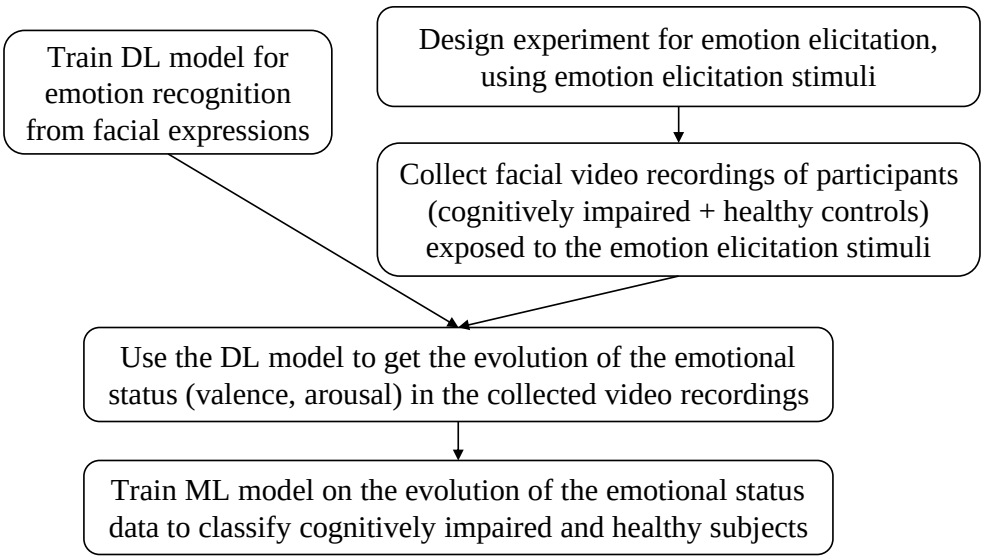

Figure 1. Proposed pipeline for classifying cognitively impaired and healthy individuals based on facial emotion analysis. A deep learning model is trained for emotion recognition and applied to videos recorded during an emotion elicitation experiment. The resulting valence and arousal patterns are used to train a machine learning model for cognitive impairment detection.

3.2. Facial Emotion Recognition Model

We propose a DL-based regression model to predict valence and arousal from face images, given the superior performance of dimensional models on emotion representation. The model is trained on a dataset with annotations in the valence–arousal space. Moreover, since the model is used in the context of a specifically designed experimental protocol eliciting spontaneous facial emotions, the training dataset is selected so as to include faces with unposed expressions (i.e., the so-called in-the-wild datasets).

3.2.1. AffectNet Dataset

AffectNet is the largest publicly available in-the-wild facial expression dataset [20], with more than 1 million facial images of individuals of various ages, sexes, and ethnicities. About half of these images (∼450k) are manually annotated for the presence of eight emotion categories (categorical model) and valence-arousal intensity (dimensional model). Since the full AffectNet database is huge (122 GB), the authors make available by default a reduced version, composed of 291,650 manually annotated images. This reduced dataset is adequate and matches our computational resources constraints; thus, it is employed in the present work. The reduced dataset is already split into the AffectNet training set (AT, 287,651 images) and AffectNet validation set (AV, 3999 images), properly representing all expression categories. The AffectNet test set has not been made publicly available.

In this work, AV is employed as our test set to evaluate the model performance, whereas AT is further split into our training and our validation set for hyperparameter tuning. The provided images are sized at 224×224 pixels (RGB color). The categorical expression label is an integer value in the range $[0, 7]$ (representing *Neutral*, *Happy*, *Sad*, *Surprise*, *Fear*, *Disgust*, *Anger*, and *Contempt* categories), while valence and arousal are provided as floating point numbers in the $[-1, 1]$ interval. The majority of the images are labeled as *Happy*, followed by *Neutral*. The remaining classes, instead, are less represented. Actually, samples frequently assume positive valence and small positive arousal values, while extreme values (especially the negative ones) seldom occur [20]. Given the current lack of large-scale emotion datasets that include individuals with cognitive impairment,

AffectNet, although based on data from the general population, is widely used in affective computing and stands as a suitable and well-supported choice for our research.

3.2.2. Model Architecture

Ngo and Yoon [32] demonstrated the effectiveness of using a deep CNN architecture, i.e., the squeeze-and-excitation network (SENet) [33] pre-trained on VGGFace2 [34] and fine-tuned on AffectNet, to predict eight categorical emotions. In this paper, an analogous transfer learning approach was employed, adapted to predict dimensional emotions instead of categorical ones.

The pre-trained SENet model available at GitHub [35] is used, with the top layers (built to perform classification only) properly modified. In detail, the last flattening and fully connected (FC) layers are replaced with a global average pooling 2D layer and two additional FC layers (with 2048 and 1024 units, respectively) to capture facial features specifically related to emotions. These are followed by three output layers with linear activation: an FC layer with eight neurons, responsible to classify categorical emotions, and two FC layers with one neuron, devoted to the prediction of valence and arousal. Even though in this paper the focus is on continuous emotional attributes, the presence of the categorical emotion output is explained in the following Section. The resulting model architecture is shown in Figure 2 and has 32,343,802 trainable parameters.

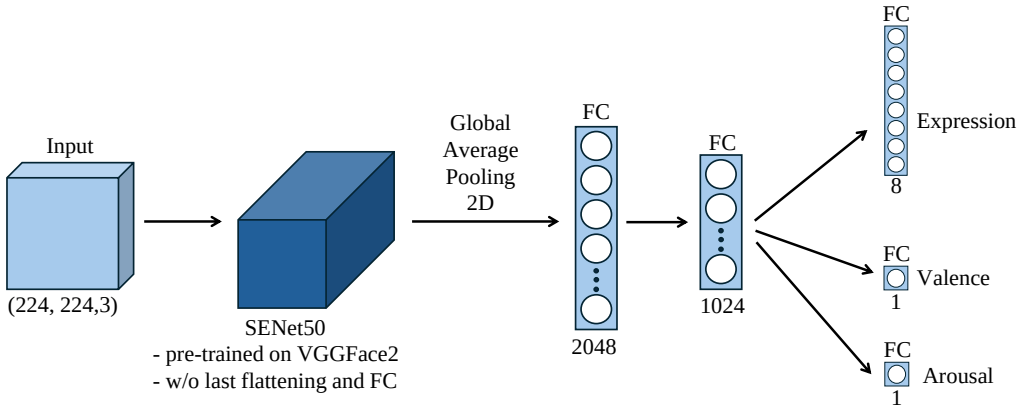

Figure 2. Architecture of the proposed CNN for valence and arousal prediction from facial images. The model is based on an SENet architecture [33], pre-trained on VGGFace2 [34] and fine-tuned on AffectNet for emotion recognition. We leverage the pre-trained SENet50 classification model available at GitHub [35], modifying the top layers to produce three outputs: valence, arousal, and categorical expression.

3.2.3. Model Training

As mentioned, our training and validation sets are extracted from the AT set. A split of 95% and 5% is implemented, achieving a training set of 273,269 images, a validation set of 14,382 images, and a test set of 3999 images. Data augmentation is applied—namely, a random horizontal flip and a random rotation in a range of 20 degrees—to avoid introducing unrealistic variations.

Although valence and arousal are generally considered independent, several studies have highlighted some dependencies between the two dimensions [36,37]; consequently, researchers have proposed systems to jointly predict multiple affect dimensions by leveraging their interdependencies. In Parthasarathy and Busso [38], the prediction of emotions is performed using speech and framed as a multi-task learning (MTL) problem, whose principal and secondary outcomes are the prediction of the target attribute (e.g., valence) and of the other attributes (e.g., arousal), respectively. This approach provided a performance improvement with respect to single-task learning, where emotional attributes are

modeled separately. Following these results, this work implements joint learning of valence and arousal with the same CNN architecture. With respect to Parthasarathy [38], we also introduce the categorical expression into the learning process, as it has been shown to improve emotion classification accuracy [23]. The categorical output is then omitted at the end of the training process. Therefore, the resulting MTL framework is trained with a weighted loss function defined as follows:

$$L = \alpha \times L_{\text{val}} + \beta \times L_{\text{aro}} + (1 - \alpha - \beta) \times L_{\text{exp}} \quad (1)$$

where L_{val} and L_{aro} are mean squared error losses related to the valence and arousal attributes, respectively, and L_{exp} is a cross-entropy loss related to the categorical expression attribute. The weights for the three attributes' losses, which must sum up to 1, are expressed using the coefficients α and β. As in [38], α and β are determined in order to maximize the performance for the target attribute. In our framework, the CNN is trained for different values of α and β; for both valence and arousal, the combination of α and β leading to the smallest error on the test set for that target attribute is selected. At the end of this process, two models are obtained, which are optimized for valence and arousal, respectively, but are trained to exploit the dependencies between valence, arousal, and the categorical expression.

In the training process, batching is used to limit the amount of memory necessary to run the network, and data shuffling is applied. The chosen batch size is 32, which is the largest possible value compliant with our computational resources constraints. Moreover, 250 steps per epoch are used; the step size establishes the number of batches to process before the epoch is considered complete. An adaptive learning rate is employed, starting from 10^{-4} and being halved after 5 epochs in which the validation loss has not improved. The training is early stopped after 10 epochs of no gain in the validation loss. The maximum number of epochs is set to 90, and the Adam optimizer is used.

3.2.4. Model Evaluation

As commonly used and formulated in [20], the performance evaluation on the test set is expressed in terms of the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2} \quad (2)$$

where N is the number of samples, $\hat{\theta}_i$ is the prediction for the ith sample, and θ_i is the ground truth of the ith sample. In addition, the concordance correlation coefficient (CCC) is computed, which is defined as

$$\text{CCC} = \frac{2\rho\sigma_{\hat{\theta}}\sigma_{\theta}}{(\mu_{\hat{\theta}} - \mu_{\theta})^2 + \sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2}; \rho = \frac{\text{COV}\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} \quad (3)$$

where ρ is the Pearson correlation coefficient, based on the covariance of the prediction ($\hat{\theta}$) and the ground truth (θ) vectors; $\mu_{\hat{\theta}}$ and μ_{θ} are their mean values; and $\sigma_{\hat{\theta}}$ and σ_{θ} their standard deviations [20]. This metric provides a measure of agreement between the ground truth and the predicted values for valence and arousal; it lies in the $[-1, 1]$ interval.

3.3. Data Collection Protocol

3.3.1. Emotion Elicitation Stimuli

The combination of visual and auditory stimuli creates a more immersive and emotionally engaging experience for participants and can potentially elicit more robust and

nuanced emotional responses compared to protocols that rely on a single modality [39]. Therefore, an emotion-eliciting video is set up, using images and sounds from IAPS (International Affective Picture System [40]) and IADS-2 (International Affective Digitized Sounds-2 [41]), respectively; these are among the most employed databases in the area of affective stimulation.

IAPS includes more than 1000 images capturing a wide array of human experiences. Each picture was rated on valence and arousal by a large group of people with diversified gender. Then, the pictures were numbered and catalogued according to the mean value and standard deviation of these affective ratings. Similarly, the IADS-2 database contains more than 100 sounds from different sources and contexts, rated analogously.

The observations are distributed across the valence–arousal plane and can be classified into five groups: high valence, high arousal (HVHA); low valence, high arousal (LVHA); low valence, low arousal (LVLA); high valence, low arousal (HVLA); and neutral. Then, 4 neutral samples and 6 samples for each of the other groups are selected, for a total of 28 images. Similar sampling techniques have been also employed in other studies on affective processing [42]. The observation selection is performed by applying filters on the valence and arousal dimensions. Moreover, to avoid potential distress, emotionally intense content from the IAPS and IADS-2 databases is excluded. Therefore, samples deemed inappropriate for our use, such as nude images or with excessively violent content, are manually discarded. The selection of audio-visual stimuli is performed in collaboration with clinicians, to prioritize safety and comfort, given the clinical vulnerability of the participants. Although this may limit the range of valence and arousal responses, the protocol ensures ethical appropriateness and is tailored to the specific needs of this sensitive population. The images are paired with sounds characterized by similar valence and arousal (identification numbers of the selected picture–sound pairs (from IAPS and IADS-2, respectively): HVHA: [8501, 367], [8185, 817], [8030, 352], [8190, 815], [8370, 363], [8492, 360]; HVLA: [5760, 811], [5000, 812], [2035, 810], [1441, 809], [2360, 151], [2530, 230]; LVHA: [9075, 260], [9410, 286], [9635.1, 292], [3530, 276], [3005.1, 296]; LVLA: [2750, 250], [9342, 382], [9280, 701], [9832, 728], [9220, 723], [7031, 708]; Neutral: [8232, 364], [1908, 170], [9422, 410], [2780, 722].). These audiovisual pairs are used to create an emotion elicitation video, whose structure is summarized in Figure 3. This protocol is inspired by that in [43].

For each subject, the experiment starts with a webcam calibration phase. The webcam is used to display the subject's image on the laptop screen for 10 s to ensure that the subject is adequately close to the screen and positioned within the desired framing.

A welcome title appears on the screen for 5.5 s to capture the subject's attention. Then, the sequence of audiovisual stimuli starts, along with the webcam recording of the subject's reactions. The 28 images are presented in a fixed, randomly determined order. A 10 s countdown is first displayed, followed by a 1 s projection of a cross in the center of the screen. Then, the image is displayed for 6 s while the paired sound is played simultaneously. The countdown in each trial alerts participants to the upcoming image, while helping them to refocus and regulate emotions; moreover, it provides a clear progression through the experiment, enhancing compliance and reducing confusion or anxiety. On the other hand, the cross fixation serves as a focal point to guide the attention to the upcoming stimulus.

At the end of the audiovisual stimuli, the recording is stopped, and a conclusion title appears for 1 s, informing the subject of the end of the experiment. The whole experiment lasts about 8 min; this duration is deemed suitable for cognitively impaired persons, minimizing cognitive load and preventing excessive fatigue or reduced engagement.

The PsychoPy software is used to set up the experiment. It allows us to define the sequence of emotional stimuli (images and sounds) and to display the emotion elicitation

video on foreground, while simultaneously having the webcam recording in a synchronized manner. The system then automatically saves the recorded video in a specified folder.

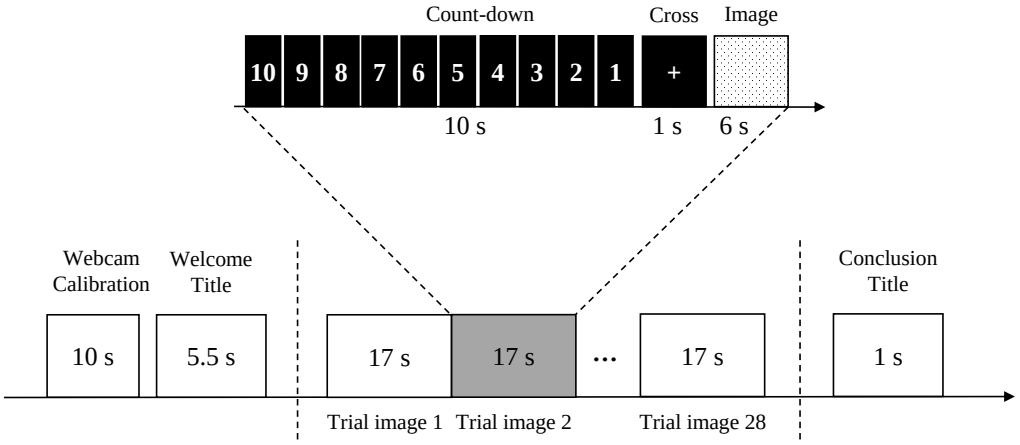

Figure 3. Structure of the emotion elicitation video protocol, inspired by [43]. The experiment begins with a webcam calibration phase, followed by a welcome title. The core of the protocol consists of 28 audiovisual trials, each comprising a 10 s countdown, a 1 s central fixation cross, and a 6 s image presentation paired with sound. A concluding title marks the end of the session. The entire procedure lasts approximately 8 min, ensuring participant engagement while minimizing cognitive load.

3.3.2. Experimental Setup

The experiments took place in a dedicated room at the Molinette Hospital—A.O.U. Città della Salute e della Scienza di Torino. The setup involved seating the subject in front of a laptop, positioned on a stable table at an appropriate height to ensure the subject's comfort. The laptop served as the primary interface for stimuli presentation and data collection. An external USB webcam (Logitech C920) was securely attached over the laptop and positioned to entirely capture the subject's facial expressions during the experiment. Adjacent to the laptop, an external Bluetooth speaker was placed to provide high-quality audio playback and to make the experiment more immersive. An illustration of our emotion elicitation setup, which is inspired by the setup of Prajapati et al. [44], is shown in Figure 4. The video recordings were performed with 1080p resolution at a frame rate of 30 fps and stored as AVI files. The adopted frame rate was considered adequate to capture facial dynamics, including micro-expressions, which may occur within time windows shorter than 200 milliseconds [45].

3.3.3. Participants

Cognitively impaired subjects were recruited among those attending the Center for Alzheimer's Disease and Related Dementias at the Department of Neuroscience and Mental Health, A.O.U. Città della Salute e della Scienza University Hospital (Turin, Italy), for early diagnosis of cognitive disorders. Patients with subjective cognitive decline were also included in this study. The participants underwent a complete neurological and neurocognitive assessment, encompassing neuropsychological tests, neuroimaging (brain magnetic resonance imaging—MRI—and positron emission tomography using 18 F-fluorodeoxyglucose—18FDG-PET), and lumbar puncture for cerebrospinal fluid biomarker analysis ($A\beta 42$, $A\beta 42/A\beta 40$, total tau, and phosphorylated tau). Based on the neurocognitive score, subjects were classified as MCI or overt dementia. The instrumental tests were used to perform differential diagnosis among the several types of CI and to rule out cases in which the condition was due to other causes. Participants were excluded if they were under the age of 18, lacked legal capacity, or presented any condition that, in the judgment of the

investigators, could hinder their ability to comply with the study protocol or compromise eligibility, such as significant motor impairments affecting facial expressiveness.

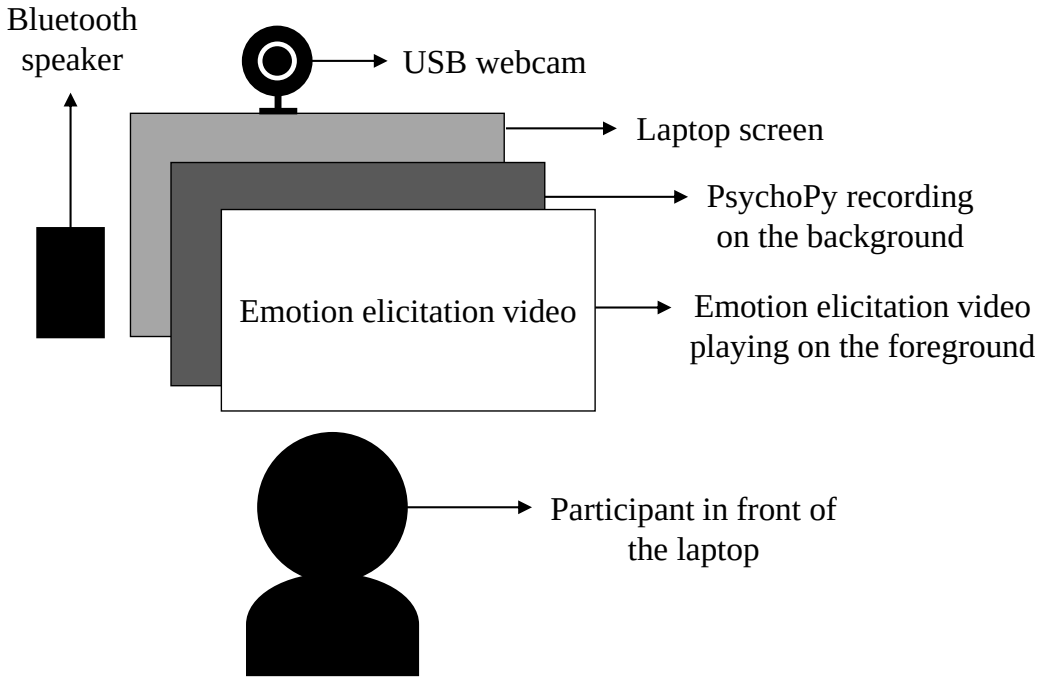

Figure 4. Illustration of the emotion elicitation setup, inspired by [44]. The participant is seated in front of a laptop that displays the emotion elicitation video in the foreground, while PsychoPy software runs in the background to manage the recording. A USB webcam mounted on top of the screen captures facial expressions throughout the session. Audio is delivered through an external Bluetooth speaker to enhance immersion and ensure high-quality playback.

As for the HC subjects, healthy volunteers between 40 and 80 years old were recruited. Exclusion criteria encompassed neurological or psychiatric disorders or other conditions preventing the execution of the experiment (e.g., blindness). HC subjects also underwent neuropsychological assessment.

The cognitive assessment was performed by an expert neuropsychologist using the MMSE, the MoCA test, the activities of daily living (ADLs), and the instrumental activities of daily living (IADLs) indices. For HC subjects, it was verified that MMSE $\geq$ 26/30, ADL $=$ 6/6 and IADL $=$ 8/8. Among participants with CI, the ones with MMSE $\geq$ 20, ADL $=$ 6/6 and IADL $\geq$ 6/8 were considered as MCI patients; instead, participants with MMSE $<$ 20 or ADL $<$ 6/6 or IADL $<$ 6/8 were considered to be affected by overt dementia.

A total of 60 individuals participated in the experiment, including 32 CI and 28 HC subjects. Among the 32 CI subjects, 23 were classified as MCI (11: likely AD; 2: mixed; 1: not specified; 9: other), and 9 were classified as overt dementia (3: AD; 2: mixed; 4: other dementia types). Demographics and relevant clinical data are summarized in Table 1.

This study was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants, and the study protocol was approved by the Ethic Committee of A.O.U. Città della Salute e della Scienza di Torino (approval number 0001863). To ensure participant privacy, data were pseudonymized through the assignment of unique codes, and all sensitive information was stored separately from associated metadata. Original video recordings were securely stored on institutional servers, with access limited to authorized personnel.

Table 1. Demographics and relevant clinical data of the participants.

	Cognitively Impaired	Healthy Controls
Number of subjects	32	28
Age (mean ± standard deviation)	69.3 ± 8.9	58.8 ± 6.9
Sex (number of females, %)	14 (43.8%)	14 (50%)
Ethnicity	Caucasian	Caucasian
Years of education (mean ± standard deviation)	12.7 ± 5.0	15.6 ± 4.8
MMSE score (mean ± standard deviation)	23.9 ± 5.3	29.2 ± 1.2
MoCA score (mean ± standard deviation)	18.7 ± 5.1	25.4 ± 2.2
Severity of cognitive impairment	23 MCI, 9 overt dementia	No cognitive impairment

3.4. Classification of Cognitively Impaired and Healthy Control Subjects

As illustrated in Figure 1, the CNN models trained for valence and arousal prediction are used to obtain the subjects' emotional states for all videos in the collected dataset. These emotional data are used to train an ML model to classify CI and HC subjects.

3.4.1. Emotional State Detection

All frames are extracted from the recorded videos (∼14k frames for each video, due to the frame rate of 30 fps). The subject's face is detected and cropped using the Holistic solution from MediaPipe [46], an open-source framework employing ML to detect face and pose landmarks in real time from video data. Face images are resized to 224 × 224 pixels and input into our trained CNN models. In this way, a value of valence and arousal is obtained for each frame. The resulting time series of valence and arousal provide insights into the evolution of the emotional state of participants during the experiment. For each subject, valence and arousal series are concatenated to create a single feature vector, used to train different classification algorithms.

3.4.2. Machine Learning Model Selection and Evaluation

Different ML algorithms are tested, using the implementation provided by scikit-learn. The algorithms deemed suitable for our classification task (also considering the limited dataset) are k-nearest neighbors (KNN), logistic regression (LR), and SVM. KNN is tuned with a grid search on the number of neighbors (3, 5, 7) and the distance metric (Euclidean, Manhattan, Chebyshev); LR is used with an L2 penalty term, the "liblinear" solver, 10^{-4} tolerance for stopping criteria, and tuned on the inverse of regularization strength C (powers of 10 from 10^{-4} to 10^{4}); SVM is used with linear kernels, tolerance 10^{-3}, and tuned on the C regularization parameter (powers of 10 from 10^{-4} to 10^{4}).

Nested cross-validation (NCV) is implemented to estimate an ML model's generalization error while simultaneously optimizing its hyperparameters. In fact, standard cross-validation (CV) is useful for mitigating test set selection bias when working with a small dataset and evaluating model performance. However, using the same CV procedure for both hyperparameter optimization and performance evaluation can lead to an overly optimistic estimate of generalization error due to overfitting.

On the other hand, NCV involves two nested CV loops. In the outer loop, model evaluation is performed through a k-CV; the dataset is repeatedly split into training ($k-1$ folds) and test sets, and the generalization error is estimated by averaging test set scores over the k splits. At each split, the $(k-1)$ folds are used to implement the inner loop, i.e., an

m-CV with a grid search for hyperparameter tuning; the data are repeatedly split into training ($m-1$ folds) and validation sets, and at each split, the best set of hyperparameters is selected based on the validation set performance. Once the hyperparameters are selected, the model is re-trained on all m folds and tested on the outer loop test set.

To ensure that folds contains approximately the same proportion of samples from each class as in the full dataset, both outer CV and inner CV are implemented with the stratified k-fold CV provided by scikit-learn. Due to the limited size of our dataset, five outer and three inner folds are employed in the NCV (k = 5, m = 3). This choice reflects standard practice, aiming to balance bias and variance by ensuring reliable performance estimates while maintaining adequate data availability within each fold. Importantly, stratification is performed at the subject level; all data belonging to a given participant are assigned entirely to either the training or validation set within each fold. This strategy prevents identity or temporal leakage across folds, ensuring a more realistic evaluation of model generalization across individuals. The process is repeated for all the ML models considered; the best model is selected as that with the highest average NCV accuracy. For this model, the optimal hyperparameter combination is adopted, i.e., those most frequently used across the outer loop folds.

4. Results

4.1. Facial Emotion Recognition

The performance of the selected CNN models on our test set for different values of α and β is shown in Table 2. The results obtained by Mollahosseini et al. [20] on the complete AffectNet dataset using a different CNN (AlexNet) are also reported as benchmarks.

According to the obtained RMSE scores, the best performance for valence is achieved with $\alpha = 0.4$, $\beta = 0.3$, while the best performance for arousal is achieved with $\alpha = 0.3$, $\beta = 0.4$. In both cases, the best results are achieved with the highest weight related to the categorical expression loss among the considered values, i.e., $1-\alpha-\beta = 0.3$ (see Equation (1)). This confirms that including the categorical expression in the training process eventually improves the model performance on valence and arousal prediction.

In terms of RMSE, the arousal dimension is generally better predicted with respect to valence. However, in terms of CCC, valence exhibits a higher concordance between predicted and true values. Our best CNN models achieve an RMSE of 0.4332 for valence prediction and 0.3641 for arousal prediction, outperforming the AffectNet benchmark on arousal prediction (RMSE = 0.402) and achieving slightly inferior performance on valence prediction (RMSE = 0.394). In terms of CCC, our system outperforms the benchmark on valence prediction (0.5571 vs. 0.541), with a comparable result on arousal prediction (0.4451 vs. 0.450). It is worth noting that the benchmark CNN is trained on a larger dataset (the complete AffectNet training set) and evaluated on the AffectNet test set, which is not publicly available. Moreover, the present work has to cope with resource constraints for CNN training, thus setting the batchsize to 32 (256 in the benchmark model), and this may have an impact on the model's performance.

4.2. Cognitive Impairment Detection

Table 3 summarizes the NCV results provided by the ML models tested for CI detection. KNN achieves the best accuracy of 76.7% in classifying cognitively impaired subjects vs. healthy controls and an F1 score of 75.4%. The optimal parameters are as follows: five neighbors and the Manhattan distance metric. In addition, the KNN classifier achieves a specificity of 92.7% and a sensitivity of 61.9%. These results highlight the model's strong ability to correctly identify HC subjects while still capturing the majority of CI

individuals, despite the inherent complexity and variability typically associated with this clinical population.

Table 2. Performance of the proposed CNN on valence and arousal prediction for different values of α and β parameters in terms of root mean square error (RMSE) and concordance correlation coefficient (CCC). The comparison with the AffectNet benchmark [20] is also reported. The best results are emphasized in bold.

	Valence		Arousal	
	RMSE	**CCC**	**RMSE**	**CCC**
AffectNet benchmark [20]	**0.394**	**0.541**	**0.402**	**0.450**
CNN, $\alpha = 0.4$, $\beta = 0.6$	0.4709	0.5013	0.391	0.3679
CNN, $\alpha = 0.5$, $\beta = 0.5$	0.4539	0.5105	0.3757	0.4213
CNN, $\alpha = 0.6$, $\beta = 0.4$	0.4661	0.4953	0.379	0.4102
CNN, $\alpha = 0.4$, $\beta = 0.5$	0.4372	0.5292	0.379	0.4067
CNN, $\alpha = 0.5$, $\beta = 0.4$	0.4567	0.5139	0.3898	0.3755
CNN, $\alpha = 0.3$, $\beta = 0.5$	0.4415	0.5377	0.3742	0.4163
CNN, $\alpha = 0.4$, $\beta = 0.4$	0.4462	0.5286	0.3827	0.3940
CNN, $\alpha = 0.5$, $\beta = 0.3$	0.4545	0.5240	0.3780	0.3947
CNN, $\alpha = 0.3$, $\beta = 0.4$	0.4360	0.5386	**0.3641**	**0.4451**
CNN, $\alpha = 0.4$, $\beta = 0.3$	**0.4332**	**0.5571**	0.3685	0.4491

Table 3. CI vs. HC classification results. The accuracy and F1 scores achieved by the tested algorithms are reported as (mean ± standard deviation), together with the corresponding optimal parameters.

	Optimal Parameter Combination	**Accuracy**	**F1 Score**
KNN	A total of 5 neighbors, Manhattan distance	**0.767 ± 0.062**	**0.754 ± 0.077**
LR	L2 penalty, tolerance = 0.0001, C = 100	0.583 ± 0.105	0.593 ± 0.102
SVM	linear kernel, tolerance = 0.001, C = 0.01	0.633 ± 0.085	0.626 ± 0.089

5. Discussion

The performance of the proposed CNN models for facial emotion recognition exceeds or matches the benchmark CNN on valence and arousal prediction, despite the latter being trained on the complete AffectNet training set and evaluated on the AffectNet test set, which is not publicly available. Overall, we can conclude that our CNN model reaches a performance that is comparable to the AffectNet benchmark [20] and is sufficient to be used within our system for cognitive impairment detection.

As for cognitive impairment detection, the achieved results demonstrate that the proposed model is promising in accurately classifying CI and HC based on the evolution of emotions and that facial expressions and the dimensional model of affect have great potential for CI detection. While still exploratory, the proposed approach could serve as a valuable complement to traditional diagnostic methods, offering a non-invasive and accessible tool to support early clinical assessment.

Comparing our model's performance with the similar research work of Fei et al. [10], our model achieves a slightly higher accuracy (76.7% vs. 73.3%). However, an exhaustive fair comparison is not possible because of the different dataset, feature engineering, neural network architecture, and CV technique. Moreover, in Fei [10], a categorical model of affect is employed.

An original aspect of the proposed work, in contrast with Fei [10], is the implementation of an emotion elicitation protocol based on standardized, widely recognized and extensively validated datasets. This is a guarantee that the protocol is grounded in reliable and well-established stimuli, enhancing its validity and generalizability. In addition, our ground truth classification of CI is based on a comprehensive set of indicators, beyond the MoCA or MMSE results as in Fei [10]. This broader approach enables the distinction between MCI and dementia, which have distinct levels of cognitive decline and functional impairment. This more precise differentiation is fundamental if we are to implement a finer-grained classification and possibly help in the differential diagnosis of different types of dementia. This is left to future developments, when more data will be available.

Limitations. It is worth recalling that the objective of this research was to set up a CI detection system based on facial emotions as potential early markers for cognitive impairment. Hence, the implementation of a more complex and performing neural network for valence and arousal prediction is outside of the scope of this paper.

A limitation of this work is the size of the collected dataset. The enrollment of individuals with CI is particularly challenging, especially in advanced stages, as it requires both valid informed consent and complete confidence that individuals can reliably engage with and complete the study protocol. A larger cohort is needed to ensure the method's robustness in view of a possible future clinical deployment. Hence, the reported accuracy (76.7%) should be interpreted as a promising preliminary result within a proof-of-concept context. As data from a larger number of subjects becomes available, AI models with higher complexity and improved generalization power can be tested.

Future work. The exploration of different model architectures for valence and arousal prediction is left to future developments, such as other advanced DL architectures showing promising results in FER, including models with attention mechanisms [25,47] and visual transformers [26,48]. Future directions also include the exploration of weighting schemes or data augmentation strategies to better account for less-represented regions in the valence–arousal space during training, which could help improve the robustness of emotional attribute prediction.

With respect to cognitive impairment detection, future work will include the collection of more data from patients with different forms and severity levels of CI in order to support a more comprehensive evaluation of the generalizability of our approach. We also plan to involve participants from more diverse ethnic and demographic backgrounds, ideally within a multicenter framework, to assess the robustness of the proposed method across different populations and clinical contexts.

In addition, future work will investigate the integration of other facial features beyond valence and arousal, as combining different types of features has proven effective in recent research [49]. The temporal dynamics of valence and arousal trajectories will also be explored, as they may provide additional informative patterns of emotional response for CI detection. Another direction for future improvements involves artefact-aware video processing techniques, as facial artefacts, such as involuntary movements, may affect emotion recognition. Our data collection was carried out in a controlled environment, which helped to limit extreme occlusions or noise; moreover, our pipeline is designed to be effective even in the presence of natural variability in facial behavior. Nevertheless, incorporating such strategies could further enhance the robustness of the proposed approach.

6. Conclusions

In this paper, a non-invasive system is described to automatically detect cognitive impairment based on facial emotion analysis and AI. A CNN model is trained on the AffectNet dataset to predict emotions from face images using a dimensional model of

affect. Then, an emotion elicitation protocol is designed to record facial expressions in response to an emotion elicitation video from IAPS and IADS-2 datasets. Facial video data of 32 CI and 28 HC subjects are collected, and the evolution of the emotional status of each subject in terms of valence and arousal is obtained. Finally, an ML model is trained on the extracted emotional responses to classify CI vs. HC. The classification algorithm achieves a cross-validation accuracy of 76.7% in distinguishing CI and HC, revealing its potential effectiveness in identifying individuals with CI by exploiting the dimensional model of affect.

Future research includes collecting more data of patients with different forms and severity levels of dementia. This would enable the exploration of multiclass classification tasks to detect not only cognitive impairment from facial emotions but also its severity and etiology, since it is known that different forms of dementia show different manifestations of facial expressions.

Author Contributions: L.B.: writing—original draft preparation, data curation, formal analysis, investigation, methodology, visualization, validation. F.L.: formal analysis, investigation, software, writing—review and editing. A.C. (Anita Coletta): formal analysis, investigation, software, writing—review and editing. G.O.: conceptualization, funding acquisition, project administration, supervision, writing—review and editing. A.C. (Aurora Cermelli): data curation, investigation, writing—review and editing. E.R.: conceptualization, funding acquisition, project administration, writing—review and editing. I.R.: conceptualization, funding acquisition, supervision, resources, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fondazione CRT, grant number 105128/2023.0366.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of A.O.U. Città della Salute e della Scienza di Torino (approval number 0001863).

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The dataset presented in this article is not readily available due to privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. World Health Organization. *Global Status Report on the Public Health Response to Dementia*; World Health Organization: Geneva, Switzerland, 2021.
2. Koyama, A.; Okereke, O.I.; Yang, T.; Blacker, D.; Selkoe, D.J.; Grodstein, F. Plasma amyloid-β as a predictor of dementia and cognitive decline: A systematic review and meta-analysis. *Arch. Neurol.* **2012**, *69*, 824–831. [CrossRef] [PubMed]
3. Alexander, G.C.; Emerson, S.; Kesselheim, A.S. Evaluation of aducanumab for Alzheimer disease: Scientific evidence and regulatory review involving efficacy, safety, and futility. *JAMA* **2021**, *325*, 1717–1718. [CrossRef]
4. van Dyck, C.H.; Swanson, C.J.; Aisen, P.; Bateman, R.J.; Chen, C.; Gee, M.; Kanekiyo, M.; Li, D.; Reyderman, L.; Cohen, S.; et al. Lecanemab in Early Alzheimer's Disease. *N. Engl. J. Med.* **2023**, *388*, 9–21. [CrossRef]
5. Jack, C.R., Jr.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Dunn, B.; Haeberlein, S.B.; Holtzman, D.M.; Jagust, W.; Jessen, F.; Karlawish, J.; et al. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's Dement.* **2018**, *14*, 535–562. [CrossRef] [PubMed]
6. Chen, K.H.; Lwi, S.J.; Hua, A.Y.; Haase, C.M.; Miller, B.L.; Levenson, R.W. Increased subjective experience of non-target emotions in patients with frontotemporal dementia and Alzheimer's disease. *Curr. Opin. Behav. Sci.* **2017**, *15*, 77–84. [CrossRef]

7. Pressman, P.S.; Chen, K.H.; Casey, J.; Sillau, S.; Chial, H.J.; Filley, C.M.; Miller, B.L.; Levenson, R.W. Incongruences between facial expression and self-reported emotional reactivity in frontotemporal dementia and related disorders. *J. Neuropsychiatry Clin. Neurosci.* **2023**, *35*, 192–201. [CrossRef] [PubMed]
8. Sun, J.; Dodge, H.H.; Mahoor, M.H. MC-ViViT: Multi-branch Classifier-ViViT to detect Mild Cognitive Impairment in older adults using facial videos. *Expert Syst. Appl.* **2024**, *238*, 121929. [CrossRef]
9. Umeda-Kameyama, Y.; Kameyama, M.; Tanaka, T.; Son, B.K.; Kojima, T.; Fukasawa, M.; Iizuka, T.; Ogawa, S.; Iijima, K.; Akishita, M. Screening of Alzheimer's disease by facial complexion using artificial intelligence. *Aging* **2021**, *13*, 1765–1772. [CrossRef]
10. Fei, Z.; Yang, E.; Yu, L.; Li, X.; Zhou, H.; Zhou, W. A Novel deep neural network-based emotion analysis system for automatic detection of mild cognitive impairment in the elderly. *Neurocomputing* **2022**, *468*, 306–316. [CrossRef]
11. Alsuhaibani, M.; Fard, A.P.; Sun, J.; Poor, F.F.; Pressman, P.S.; Mahoor, M.H. A Review of Deep Learning Approaches for Non-Invasive Cognitive Impairment Detection. *arXiv* **2024**. Available online: http://arxiv.org/abs/2410.19898 (accessed on 29 January 2025). [CrossRef]
12. Dodge, H.H.; Yu, K.; Wu, C.Y.; Pruitt, P.J.; Asgari, M.; Kaye, J.A.; Hampstead, B.M.; Struble, L.; Potempa, K.; Lichtenberg, P.; et al. Internet-Based Conversational Engagement Randomized Controlled Clinical Trial (I-CONECT) Among Socially Isolated Adults 75+ Years Old With Normal Cognition or Mild Cognitive Impairment: Topline Results. *Gerontol.* **2023**, *64*, gnad147. [CrossRef]
13. Zheng, C.; Bouazizi, M.; Ohtsuki, T.; Kitazawa, M.; Horigome, T.; Kishimoto, T. Detecting Dementia from Face-Related Features with Automated Computational Methods. *Bioengineering* **2023**, *10*, 862. [CrossRef]
14. Kishimoto, T.; Takamiya, A.; Liang, K.; Funaki, K.; Fujita, T.; Kitazawa, M.; Yoshimura, M.; Tazawa, Y.; Horigome, T.; Eguchi, Y.; et al. The project for objective measures using computational psychiatry technology (PROMPT): Rationale, design, and methodology. *Contemp. Clin. Trials Commun.* **2020**, *19*, 100649. [CrossRef]
15. Jiang, Z.; Seyedi, S.; Haque, R.U.; Pongos, A.L.; Vickers, K.L.; Manzanares, C.M.; Lah, J.J.; Levey, A.I.; Clifford, G.D. Automated analysis of facial emotions in subjects with cognitive impairment. *PLoS ONE* **2022**, *17*, e0262527. [CrossRef]
16. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef] [PubMed]
17. Plutchik, R. A psycho evolutionary theory of emotions. *Soc. Sci. Inf.* **1982**, *21*, 529–553. [CrossRef]
18. Ekman, P.; Friesen, W.V. Facial action coding system. *Environ. Psychol. Nonverbal Behav.* **1978**. [CrossRef]
19. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
20. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
21. Schoneveld, L.; Othmani, A. Towards a General Deep Feature Extractor for Facial Expression Recognition. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2339–2342. [CrossRef]
22. Singh, S.; Nasoz, F. Facial Expression Recognition with Convolutional Neural Networks. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 0324–0328. [CrossRef]
23. Handrich, S.; Dinges, L.; Al-Hamadi, A.; Werner, P.; Al Aghbari, Z. Simultaneous prediction of valence/arousal and emotions on AffectNet, Aff-Wild and AFEW-VA. *Procedia Comput. Sci.* **2020**, *170*, 634–641. [CrossRef]
24. Teixeira, T.; Granger, E.; Lameiras Koerich, A. Continuous Emotion Recognition with Spatiotemporal Convolutional Neural Networks. *Appl. Sci.* **2021**, *11*, 11738. [CrossRef]
25. Li, J.; Jin, K.; Zhou, D.; Kubota, N.; Ju, Z. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **2020**, *411*, 340–350. [CrossRef]
26. Huang, Q.; Huang, C.; Wang, X.; Jiang, F. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [CrossRef]
27. Xiaohua, W.; Muzi, P.; Lijuan, P.; Min, H.; Chunhua, J.; Fuji, R. Two-level attention with two-stage multi-task learning for facial emotion recognition. *J. Vis. Commun. Image Represent.* **2019**, *62*, 217–225. [CrossRef]
28. Karnati, M.; Seal, A.; Bhattacharjee, D.; Yazidi, A.; Krejcar, O. Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–31. [CrossRef]
29. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 29 January 2025).
30. Peirce, J.; Gray, J.R.; Simpson, S.; MacAskill, M.; Höchenberger, R.; Sogo, H.; Kastman, E.; Lindeløv, J.K. PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **2019**, *51*, 195–203. [CrossRef] [PubMed]
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
32. Ngo, Q.; Yoon, S. Facial Expression Recognition Based on Weighted-Cluster Loss and Deep Transfer Learning Using a Highly Imbalanced Dataset. *Sensors* **2020**, *20*, 2639. [CrossRef] [PubMed]

33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
34. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74. [CrossRef]
35. Keras VGGFace. VGGFace Implementation with Keras Framework. Available online: https://github.com/rcmalli/keras-vggface (accessed on 29 January 2025).
36. Yik, M.; Mues, C.; Sze, I.N.; Kuppens, P.; Tuerlinckx, F.; De Roover, K.; Kwok, F.H.; Schwartz, S.H.; Abu-Hilal, M.; Adebayo, D.F.; et al. On the relationship between valence and arousal in samples across the globe. *Emotion* **2023**, *23*, 332–344. [CrossRef]
37. Nandy, R.; Nandy, K.; Walters, S.T. Relationship between valence and arousal for subjective experience in a real-life setting for supportive housing residents: Results from an ecological momentary assessment study. *JMIR Form. Res.* **2023**, *7*, e34989. [CrossRef]
38. Parthasarathy, S.; Busso, C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1103–1107. [CrossRef]
39. Horvat, M.; Kukolja, D.; Ivanec, D. Comparing affective responses to standardized pictures and videos: A study report. In Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 25–29 May 2015; pp. 1394–1398. [CrossRef]
40. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual*; Technical Report Technical Report A-8; University of Florida, NIMH Center for the Study of Emotion and Attention: Gainesville, FL, USA, 2008.
41. Bradley, M.M.; Lang, P.J. *The International Affective Digitized Sounds (IADS-2): Affective Ratings of Sounds and Instruction Manual*; Technical Report Technical Report B-3; University of Florida, NIMH Center for the Study of Emotion and Attention: Gainesville, FL, USA, 2007.
42. Yuen, K.; Johnston, S.; Martino, F.; Sorger, B.; Formisano, E.; Linden, D.; Goebel, R. Pattern classification predicts individuals' responses to affective stimuli. *Transl. Neurosci.* **2012**, *3*, 278–287. [CrossRef]
43. Petrantonakis, P.C.; Hadjileontiadis, L.J. Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Trans. Affect. Comput.* **2010**, *1*, 81–97. [CrossRef]
44. Prajapati, V.; Guha, R.; Routray, A. Multimodal prediction of trait emotional intelligence-Through affective changes measured using non-contact based physiological measures. *PLoS ONE* **2021**, *16*, e0254335. [CrossRef] [PubMed]
45. Merghani, W.; Davison, A.K.; Yap, M.H. A Review on Facial Micro-Expressions Analysis: Datasets, Features and Metrics. *arXiv* **2018**. Available online: http://arxiv.org/abs/1805.02397 (accessed on 29 January 2025). [CrossRef]
46. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv* **2019**. Available online: http://arxiv.org/abs/1906.08172 (accessed on 29 January 2025). [CrossRef]
47. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition. *Biomimetics* **2023**, *8*, 199. [CrossRef]
48. Ma, F.; Sun, B.; Li, S. Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1236–1248. [CrossRef]
49. Okunishi, T.; Zheng, C.; Bouazizi, M.; Ohtsuki, T.; Kitazawa, M.; Horigome, T.; Kishimoto, T. Dementia and MCI Detection Based on Comprehensive Facial Expression Analysis From Videos During Conversation. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 3537–3548. [CrossRef]