Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (38^{th}cycle)

Deep learning and computer vision approaches for clinical support: from cognitive impairment detection to pediatric care

By

Letizia Bergamasco

******

Supervisor(s):
Prof. Gabriella Olmo, Supervisor
Claudio Gianni Pastrone, Co-Supervisor
Marco Gavelli, Co-Supervisor

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Letizia Bergamasco
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Acknowledgements

I would like to express my deepest gratitude to my Supervisor, Prof. Gabriella Olmo, for her continuous support, inspiring guidance, and encouragement throughout this Ph.D. journey.

I am grateful to Politecnico di Torino and to my colleagues from the *Data Analytics and Technologies for Health Lab* at the *Department of Control and Computer Engineering* for the stimulating academic environment and the exchange of ideas that accompanied this research.

This Ph.D. research was funded by LINKS Foundation, whose support is gratefully acknowledged. I am especially thankful to Claudio Gianni Pastrone and Marco Gavelli for their involvement as Co-Supervisors, and to my colleagues from the *Connected Systems and Cybersecurity* research domain at LINKS Foundation for providing a supportive and collaborative environment during these years.

I am also grateful to my collaborators and co-authors from the partner institutions involved in these research activities. The interdisciplinary collaboration with medical professionals has been an inspiring and professionally enriching experience.

Finally, I wish to thank my family and friends for their support throughout this journey. Their presence and encouragement have been a constant source of motivation.

Abstract

This PhD research explores the application of deep learning and computer vision to support clinical decision-making in cognitive impairment detection and pediatric care. Central to this approach is the analysis of camera-based data, which provides a non-invasive and accessible way to capture subtle behavioral and emotional signals relevant for clinical assessment.

Dementia is the most common neurodegenerative disorder and a major cause of disability in the elderly. Although no definitive cure is currently available, emerging treatments show promise if administered at a very early stage. Early detection is therefore essential, both to improve patients' quality of life and to enable their participation in clinical studies. In individuals with cognitive impairment, facial expressions are often altered, with distinct patterns across dementia types. This research investigates facial expression analysis as a potential non-invasive biomarker to support early detection and differential diagnosis. An artificial intelligence–based system is proposed to detect cognitive impairment through facial emotion analysis, achieving 76.0% accuracy in distinguishing individuals with mild cognitive impairment from healthy controls and 75.4% accuracy in discriminating Alzheimer's disease from other forms of cognitive decline. These findings highlight the promise of the proposed approach as a supportive tool for early and differential diagnosis of dementia, moving toward non-invasive screening methods.

Beyond cognitive impairment, this research also extends to pediatric care, focusing on the crucial challenge of automatic and contactless pain assessment in newborns and young children. Infants and non-verbal children are often unable to express their discomfort clearly, making it challenging for healthcare professionals to accurately assess and respond to their pain. Traditional methods, based on subjective observation and scoring systems, can be inconsistent and time-consuming. An artificial intelligence-based approach analyzing facial expressions, body movements, and

other visual cues offers a more objective solution, reducing medical staff's burden while improving the quality of care. In this research, a camera-based system is proposed for the automatic detection of behavioral parameters, and the feasibility of its usage in the Pediatric Emergency Department is evaluated. Additionally, a deep learning framework is introduced for newborn pain assessment through facial expression analysis in the Neonatology Department, reaching 88.8% accuracy in pain detection. Results confirm the feasibility of automated pain detection in real-world conditions, including scenarios where pain management strategies, such as pacifier use, are applied. By embedding explainable artificial intelligence, model transparency and trust among healthcare professionals are strengthened. This work lays the foundation for automated systems capable of integrating, standardizing, and enhancing human pain assessment in clinical practice.

Recently, Large Language Models have emerged as promising tools to support pediatric care. Yet, their effectiveness in assisting the diagnostic process within the Pediatric Emergency Department remains largely unexplored. To address this gap, the diagnostic performance of five state-of-the-art Large Language Models is evaluated against physicians on real pediatric emergency cases of varying complexity. Results show that advanced models can achieve diagnostic accuracy comparable to, and in some cases surpassing, that of experienced clinicians. These findings underscore the potential of Large Language Models as supportive assistants in pediatric emergency care, enhancing clinical decision-making without replacing physicians' judgment, while emphasizing the need for structured protocols to ensure their safe and effective integration into practice.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Overview and Motivation

In recent years, deep learning and computer vision have opened up new possibilities for supporting clinical work in a way that is increasingly accessible, efficient, and non-invasive. These technologies make it possible to analyze complex patterns in images and videos, such as facial expressions or body movements, that often carry valuable information about a person's health status. In healthcare, where timely and accurate decisions are crucial, having tools that can assist in interpreting subtle visual cues may significantly improve diagnosis and monitoring. This potential makes these technologies a valuable ally in clinical settings, not as a substitute for human observation or medical expertise, but as a complementary tool to support more objective and informed decision-making.

In this context, this thesis presents the research carried out by the author during her PhD program. The work brings together different studies, all aimed at leveraging deep learning and computer vision techniques to support and improve clinical practice. The research has been developed through a close collaboration between Politecnico di Torino, LINKS Foundation, and several healthcare institutions in Turin, Italy. It originates from concrete needs, observations, and research questions raised by clinicians during everyday medical practice, and has been shaped through continuous dialogue with the healthcare professionals involved. A key aspect of this work is its multidisciplinary nature, which integrates the perspectives and expertise

of engineers, physicians, and clinical staff in a shared effort to develop meaningful and applicable technological solutions.

In particular, the two main focus areas of this research are cognitive impairment detection and pediatric care. These domains have been selected not only for their clinical relevance, but also for the unmet needs that emerged from direct interaction with healthcare professionals.

In the case of cognitive disorders, and dementia in particular, early detection remains a major challenge. Dementia is one of the most common causes of disability and dependency among older adults, with significant consequences for patients, families, and healthcare systems. While no curative treatments are currently available, new pharmacological and rehabilitative strategies are under clinical evaluation, and initial results suggest that their effectiveness may depend on administration during the very early, even preclinical, stages of the disease. This makes the development of non-invasive, widely accessible screening tools a critical objective. Facial expressions are known to be altered in individuals with cognitive decline and to vary across different forms of dementia; therefore, they offer a promising avenue for exploration. Their analysis, largely overlooked in standard diagnostic procedures, may contribute both to early detection and to more accurate differential diagnosis.

The second area of investigation is pediatric care, with a particular focus on the automatic assessment of pain in newborns and non-verbal children. In these patients, pain measurement often relies on the subjective interpretation of behavioral cues by caregivers and medical staff; existing assessment approaches are time-consuming and prone to inter-observer variability. Developing contactless, artificial intelligence-based systems capable of interpreting facial and body signals in an objective and efficient manner could greatly support clinicians in delivering timely and appropriate care, while reducing the cognitive load on healthcare personnel and improving the quality of life of young patients.

Alongside the exploration of vision-based approaches, the thesis also considers the role of Large Language Models (LLMs) in supporting clinical decision-making for pediatric care. Unlike computer vision systems, which analyze visual cues, LLMs operate on textual and clinical case data, offering complementary support to physicians in diagnostic reasoning. Their potential becomes particularly relevant in pediatric emergency care, where rapid and reliable decision-making is crucial. However, since their role as a diagnostic support tool in the Pediatric Emergency

Department (PED) has not been explored yet, studies to evaluate their diagnostic efficacy are necessary.

1.2 Scientific Contributions

The methods and results of the research activities presented in this thesis have been published within several scientific contributions, including three journal papers and two conference papers, respectively:

- **Bergamasco, L.**; Lorenzo, F.; Coletta, A.; Olmo, G.; Cermelli, A.; Rubino, E.; Rainero, I. (2025). Automatic Detection of Cognitive Impairment through Facial Emotion Analysis. *Applied Sciences*, 15(16), 9103.
- **Bergamasco, L.**; Coletta, A.; Olmo, G.; Cermelli, A.; Rubino, E.; Rainero, I. (2025). AI-based Facial Emotion Analysis for Early and Differential Diagnosis of Dementia. *Bioengineering*, 12(10), 1082.
- Del Monte, F.; Barolo, R.; Circhetta, M.; Delmonaco, A. G.; Castagno, E.; Pivetta, E.; **Bergamasco, L.**; Franco, M.; Olmo, G.; Bondone, C. (2025). Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study. *Frontiers in Digital Health*, 7, 1624786.
- **Bergamasco, L.**; Gavelli, M.; Fadda, C.; Parodi, E.; Bondone, C.; Castagno, E. (2023). Measurement of Acute Pain in the Pediatric Emergency Department Through Automatic Detection of Behavioral Parameters: A Pilot Study. In *International Work-Conference on Bioinformatics and Biomedical Engineering* (pp. 469-481). Springer Nature Switzerland.
- **Bergamasco, L.**; Lattanzi, M.; Gavelli, M.; Pastrone, C.; Olmo, G.; Borsotti, L.; Parodi, E. (2024). Pain Assessment in Neonatal Clinical Practice via Facial Expression Analysis and Deep Learning. In *International Work-Conference on Bioinformatics and Biomedical Engineering* (pp. 249-263). Springer Nature Switzerland.

The following publications were also produced during the PhD period, although they are not directly related to the specific topics addressed in this thesis:

- Bertozzi, N.; Geraci, A.; **Bergamasco, L.**; Ferrera, E.; Pristeri, E.; Pastrone, C. (2025). A Distributed Event-Orchestrated Digital Twin Architecture for Optimizing Energy-Intensive Industries. In *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security* (pp. 337-344). SciTePress.

- Alberti, E.; Alvarez-Napagao, S.; Anaya, V.; Barroso, M.; Barrué, C.; Beecks, C.; **Bergamasco, L.**; Chala, S. A.; Gimenez-Abalos, V.; Graß, A.; Hinjos, D.; Holtkemper, M.; Jakubiak, N.; Nizamis, A.; Pristeri, E.; Sànchez-Marrè, M.; Schlake, G.; Scholz, J.; Scivoletto, G.; Walter, S. (2024). AI Lifecycle Zero-Touch Orchestration within the Edge-to-Cloud Continuum for Industry 5.0. *Systems*, 12(2), 48.

- Viviani, P.; Gesmundo, I.; Ghinato, E.; Agudelo-Toro, A.; Vercellino, C.; Vitali, G.; **Bergamasco, L.**; Scionti, A.; Ghislieri, M.; Agostini, V.; Terzo, O.; Scherberger, H. (2023). Deep Learning for real-time neural decoding of grasp. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 379-393). Springer Nature Switzerland.

- Morra, L.; Azzari, A.; **Bergamasco, L.**; Braga, M.; Capogrosso, L.; Delrio, F.; Di Giacomo, G.; Eiraudo, S.; Ghione, G.; Giudice, R.; Koudounas, A.; Piano, L.; Rege Cambrin, D.; Risso, M.; Rondina, M.; Russo, A. S.; Russo, M.; Taioli, F.; Vaiani, L.; Vercellino, C. (2023). Designing Logic Tensor Networks for Visual Sudoku Puzzle Classification. In *17th International Workshop on Neural-Symbolic Learning and Reasoning* (pp. 223-232).

1.3 Thesis Outline

The remainder of this thesis is structured around a series of research studies, each addressing a distinct topic within the broader framework of AI-based clinical support. Each Chapter is dedicated to a specific research topic and follows a consistent structure: it introduces the clinical problem, reviews relevant background and related work, describes the adopted methodology and materials, presents the results, discusses their implications, and ends with concluding remarks and perspectives for future work.

- Chapter 2 focuses on the use of facial expression analysis for the early and differential diagnosis of dementia, exploring its potential as a non-invasive biomarker. This work has been carried out in collaboration with Dipartimento di Neuroscienze - Università degli Studi di Torino, S.C. Neurologia 1, and Centro di Ricerca Clinica Cefalee - Dipartimento di Neuroscienze e Salute Mentale - A.O.U. Città della Salute e della Scienza di Torino.

- Chapter 3 presents a research work on the automatic assessment of pain in infants in the PED, with an emphasis on contactless and objective evaluation methods based on visual cues. It has involved a collaboration with S.C. Pediatria e Neonatologia - A.O. Ordine Mauriziano di Torino, S.C. Pediatria d'Urgenza - Ospedale Infantile Regina Margherita, and Scuola di Specializzazione in Pediatria - Università degli Studi di Torino.

- Chapter 4 presents research on the automatic assessment of pain in newborns, based on facial expression analysis and deep learning. This work has resulted from a collaboration with S.C. Pediatria e Neonatologia - A.O. Ordine Mauriziano di Torino.

- Chapter 5 explores the application of LLMs for diagnostic support in the PED. This study has involved the collaboration of S.C. Pediatria d'Urgenza - Ospedale Infantile Regina Margherita, Scuola di Specializzazione in Pediatria - Università degli Studi di Torino, and Dipartimento di Scienze Cliniche e Biologiche - Università degli Studi di Torino.

- Chapter 6 presents the overall conclusions of the thesis, summarizing the main findings, discussing the challenges and limitations encountered, and outlining future perspectives.

Chapter 2

Facial Expression Analysis for Early and Differential Diagnosis of Dementia

This work has been carried out in collaboration with Dipartimento di Neuroscienze - Università degli Studi di Torino, S.C. Neurologia 1, and Centro di Ricerca Clinica Cefalee - Dipartimento di Neuroscienze e Salute Mentale - A.O.U. Città della Salute e della Scienza di Torino. Part of the work described in this Chapter has been published in two papers: Bergamasco et al. (2025), *Automatic Detection of Cognitive Impairment through Facial Emotion Analysis* [1], and Bergamasco et al. (2025), *AI-based Facial Emotion Analysis for Early and Differential Diagnosis of Dementia* [2]. This research was funded by Fondazione CRT, grant number 105128/2023.0366.

2.1 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 syndromes in modern society at a global level [3].

The most common form is Alzheimer's disease (AD), typically presenting with short-term memory impairment at the onset [4]. Less frequent types of dementia have a relevant social and clinical impact, and their differential diagnosis is often difficult; these include vascular dementia (VD) [5], frontotemporal dementia (FTD) [6], dementia with Lewy bodies (DLB) [7], and mixed forms.

Current therapies focus on providing temporary symptom relief but have little to no effectiveness in modifying disease progression [8]. However, the Food and Drug Administration has recently approved disease-modifying therapies for AD [9, 10], 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 ($A\beta 42$, $A\beta 42/A\beta 40$, total tau, and phosphorylated tau 181) [11]. 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 [12, 13]; 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 [14–16].

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. For these purposes, an emotion elicitation protocol is set up and video recordings are collected from subjects whose clinical diagnosis are supported by relevant biomarkers, including AD-specific biomarkers in the cerebrospinal fluid. The obtained results demonstrate that the proposed system achieves good performance not only in detecting CI, but also in identifying MCI patients and discriminating AD, thus showing promising support for the early and differential diagnosis of dementia.

The main contributions are summarized as follows:

- The study addresses the important problem of detecting cognitive impairment from its very beginning (hence, including MCI) using non-invasive and cost-effective techniques.
- The assumption that facial expressions in response to emotional elicitation are different in cognitively impaired and healthy subjects is validated.
- To this end, an AI framework based on facial emotion data using a dimensional model of affect is proposed. An emotion elicitation protocol based on standardized stimuli is designed and tested.
- The system is evaluated 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.
- The capability of the proposed system to discriminate different stages of CI (MCI, overt dementia), and differentiate AD from other forms of cognitive impairment is assessed. To the best of our knowledge, this is the first study to propose an automated method to differentiate between diverse etiologies of dementia based on facial emotion analysis.

2.2 Background

This Section presents related works on DL-based cognitive impairment detection using different types of facial features (Section 2.2.1). Moreover, it addresses in more detail the task of automated facial emotion recognition, reviewing the main models used to represent human emotions, and popular DL techniques and datasets in literature (Section 2.2.2).

2.2.1 Deep Learning for Cognitive Impairment Detection Using Facial Features

A recent review paper by Alsuhaibani et al. [17] 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. [14] worked on the I-CONECT dataset [18], 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. [15] 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. [19] used face mesh, histograms of oriented gradients (HOGs), and action units (AUs—see Section 2.2.2 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 [20]. This dataset encompasses 447 videos of 117 subjects, including HC and various pathological individuals affected by dementia, bipolar

disorder, and depression. However, Zheng [19] 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. These facial features are of particular interest, not only for exhibiting good performance, but also for facilitating model interpretability for clinicians, since emotional regulation is directly affected by dementia. To show that cognitively impaired subjects express facial emotions differently from cognitively unimpaired ones, Jiang et al. [21] 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. [16] 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 in terms of automatic CI detection from facial emotions, Fei [16] 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.

The above-mentioned studies focused specifically on the detection of MCI or dementia, or targeted CI detection by grouping subjects with MCI and dementia. However, very few works in the literature tackled the differentiation of different stages of the disease combined (e.g., MCI and overt dementia), or different underlying etiologies (e.g., AD and other forms of neurodegenerative conditions). A recent work by Okunishi et al. [22] proposed a method to detect MCI and dementia based on AUs, eight emotion categories, valence-arousal, and face embeddings. By extracting and combining all these features from video recordings, they achieved an 86.2% accuracy for dementia detection and an 83.4% accuracy for MCI detection, on a selected subset of the PROMPT dataset [20]. Chu et al. [23] recruited 95 participants (41 MCI, 54 mild to moderate dementia) and recorded them while administered the Short Portable Mental Status Questionnaire. They performed a binary classification

of MCI and dementia with DL models trained on visual and speech features, reaching a 76.0% accuracy (rising to 88% when excluding depression and anxiety). On the other hand, Jiang et al. [21] successfully differentiated individuals with CI from HC; however, they were unable to differentiate among the underlying etiologies.

2.2.2 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 [24] 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 [25] (Figure 2.1), 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.

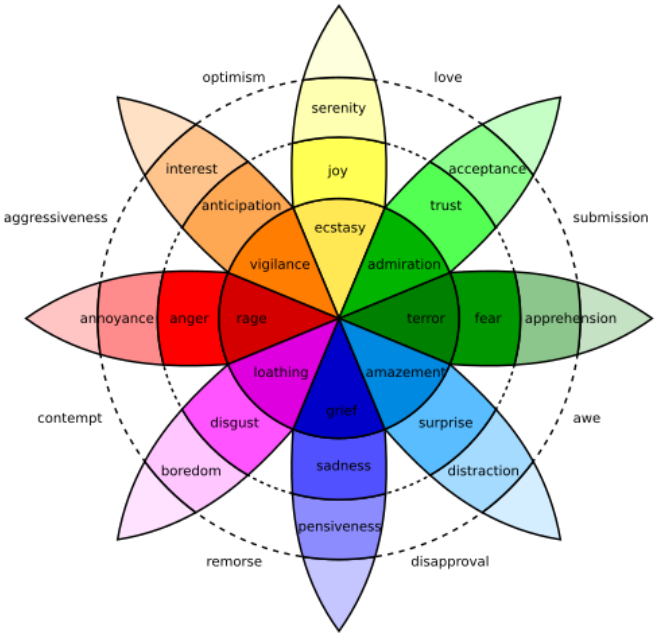


Fig. 2.1 Robert Plutchik's Wheel of Emotions [26].

The facial action coding system (FACS) [27] 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 (Table 2.1). 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.

Table 2.1 FACS codes of the six basic emotions (adapted from [28, 29]).

Emotion	**Action Units (AUs)**	**Muscle Description**
Happiness	6+12	Cheek Raiser (6), Lip Corner Puller (12)
Sadness	1+4+15	Inner Brow Raiser (1), Brow Lowerer (4), Lip Corner Depressor (15)
Surprise	1+2+5+26	Inner Brow Raiser (1), Outer Brow Raiser (2), Upper Lid Raiser (5), Jaw Drop (26)
Fear	1+2+4+5+7+20+26	Inner Brow Raiser (1), Outer Brow Raiser (2), Brow Lowerer (4), Upper Lid Raiser (5), Lid Tightener (7), Lip Stretcher (20), Jaw Drop (26)
Anger	4+5+7+23	Brow Lowerer (4), Upper Lid Raiser (5), Lid Tightener (7), Lip Tightener (23)
Disgust	9+15+16	Nose Wrinkler (9), Lip Corner Depressor (15), Lower Lip Depressor (16)

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 [30] (Figure 2.2) 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).

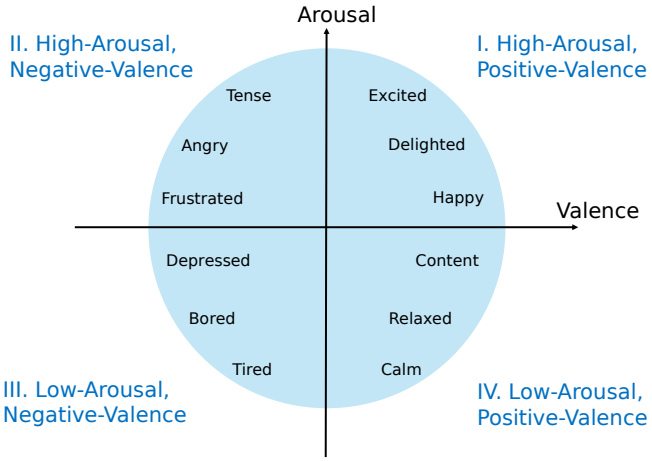


Fig. 2.2 The circumplex model of affect (inspired by [30]).

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 [31]. Table 2.2 reports a summary of literature datasets with affect annotations.

Table 2.2 Literature datasets with affect annotations.

Dataset	**Affect Model**	**Condition**	**Details**	**Availability**
AffectNet [31]	8 emotion categories, valence-arousal	in-the-wild	∼420 000 images manually annotated, ∼550 000 images automatically annotated	request form
CK+ [32]	7 emotion categories, 30 AUs	controlled posed	∼500 image sequences from 100 subjects	request form
FER-2013 [33]	7 emotion categories	in-the-wild	∼35 000 images queried from web	available
AFEW-VA [34]	valence-arousal	in-the-wild	600 videos, ∼30 000 frames	available
JAFFE [35]	7 emotion categories	controlled posed	213 images from 10 subjects	request form
Aff-Wild2 [36]	valence-arousal	in-the-wild	260 videos, ∼1 500 000 frames	request form
RECOLA [37]	valence-arousal, self assessment	controlled spontaneous	multi-modal audio, video, ECG and EDA from 46 subjects	request form
OMG-Emotion [38]	7 emotion categories, valence-arousal	in-the-wild	567 1-minute videos	available

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 [39, 40] or seldom dimensional models [31, 41, 42]. Moreover, numerous recent studies have made use of the attention mechanism, which proved to be effective in FER [43, 44]. Most papers have applied the attention mechanism to CNNs in discrete emotion classification, but Xiaohua et al. [45] 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 [46], 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.

2.3 Materials and Methods

This Section starts with an overview of the proposed system for cognitive impairment detection (Section 2.3.1). Then, all the details of the facial emotion recognition model employed in this work are presented, from model architecture to training and evaluation methods (Section 2.3.2). Afterwards, the data collection protocol employed in this work is introduced, covering the emotion elicitation stimuli, the experimental setup, the recruitment of participants, and the description of the collected dataset (Section 2.3.3). At the end, the cognitive impairment detection process is described, including the definition of the machine learning tasks and models used (Section 2.3.4).

2.3.1 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 [31]. (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 2.3.

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 [47] 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 [48]. The ML classifiers for CI detection are realized using the scikit-learn Python library [49].

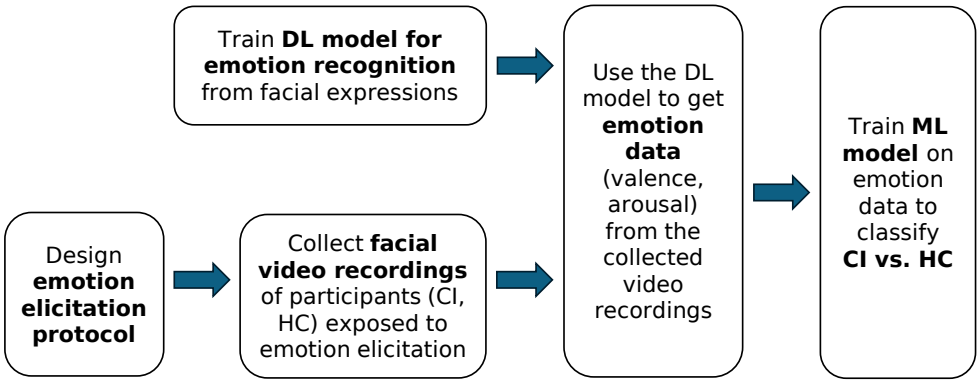


Fig. 2.3 Proposed pipeline for classifying cognitively impaired and healthy individuals based on facial emotion analysis.

2.3.2 Facial Emotion Recognition Model

In this work a DL-based regression model to predict valence and arousal from face images is proposed, 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).

AffectNet Dataset

AffectNet is the largest publicly available in-the-wild facial expression dataset [31], with more than 1 million facial images of individuals of various ages, sexes, and ethnicities. About half of these images ($\sim$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 (Figure 2.4). Actually, samples frequently assume positive valence and small positive arousal values, while extreme values (especially the negative ones) seldom occur [31] (Figure 2.5). Since in a regression task data imbalance is generally considered less critical than in classification tasks, no data balancing techniques are applied during model training, while acknowledging the potential challenges associated with this choice. This same approach was also followed by the authors of AffectNet in building their dimensional model [31]. 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 this research.

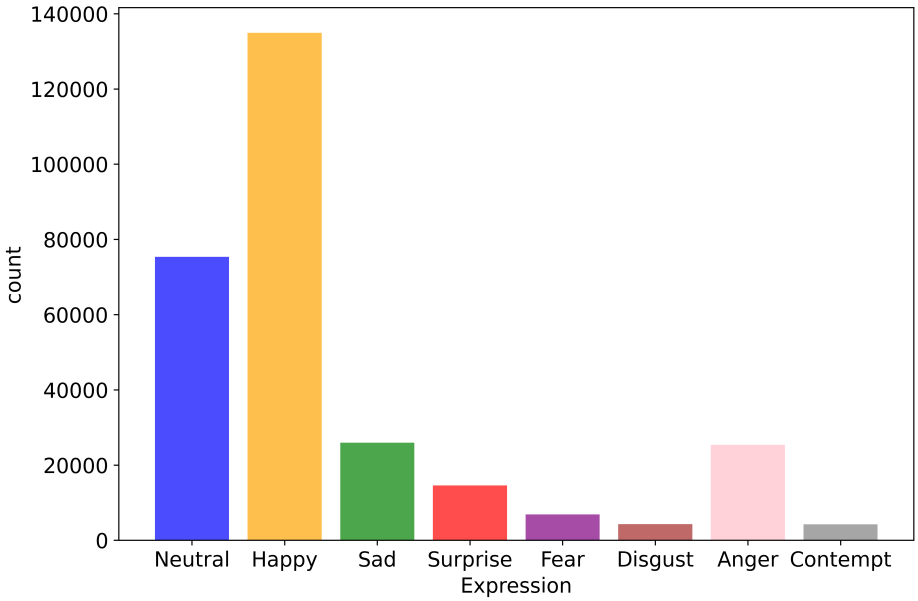


Fig. 2.4 Distribution of emotion categories in AffectNet (AT + AV).

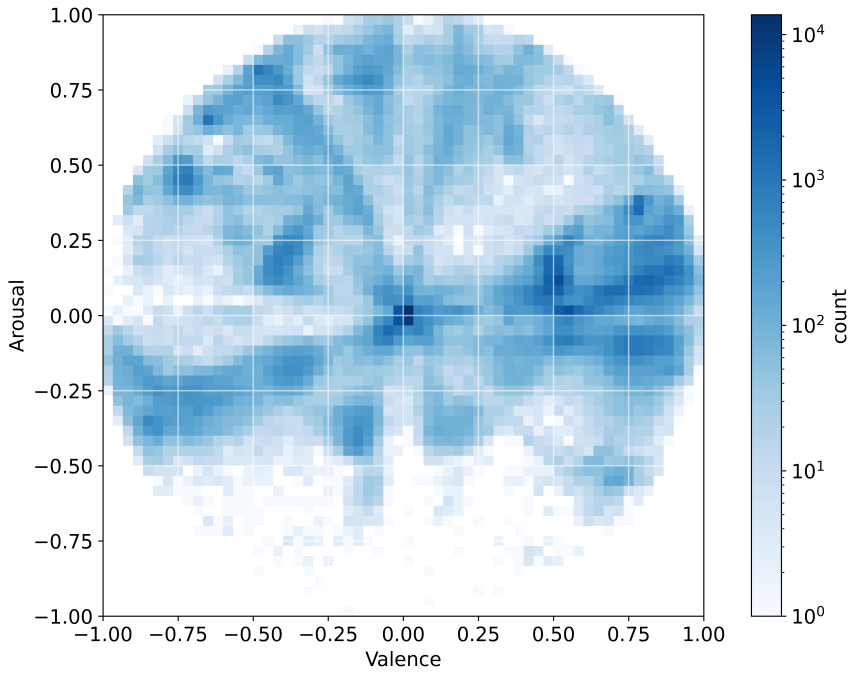


Fig. 2.5 Distribution of valence and arousal in AffectNet (AT + AV).

Model Architecture

In [50], Ngo and Yoon demonstrated the effectiveness of using a deep CNN architecture, i.e., the squeeze-and-excitation network (SENet) [51] pre-trained on VGGFace2 [52] and fine-tuned on AffectNet, to predict eight categorical emotions. In this work, an analogous transfer learning approach is employed, adapted to predict dimensional emotions instead of categorical ones.

The pre-trained SENet model available at GitHub [53] 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 work 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 2.6 and has 32,343,802 trainable parameters.

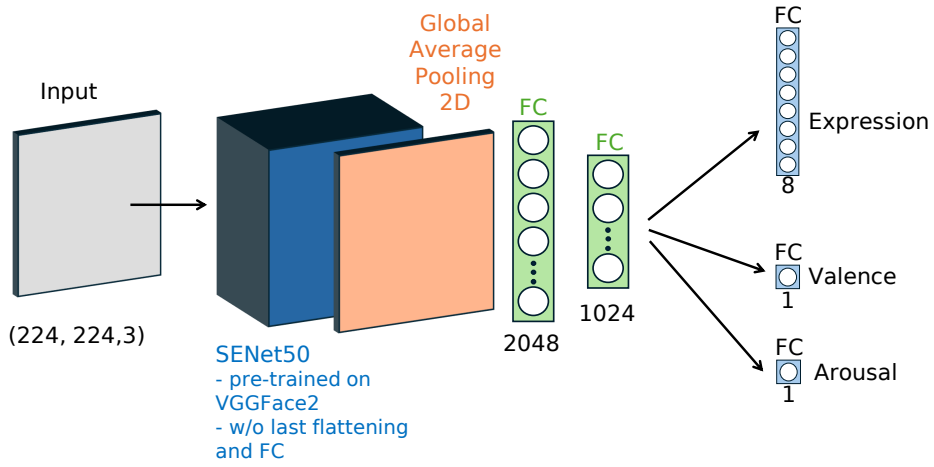


Fig. 2.6 Architecture of the proposed CNN for valence and arousal prediction from facial images (adapted from [1]).

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 [54, 55]; consequently, researchers have proposed systems to jointly predict multiple affect dimensions by leveraging their interdependencies. In Parthasarathy and Busso [56], 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 [56], the categorical expression is here introduced into the learning process, as it has been shown to improve emotion classification accuracy [41]. 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:

$$L = \alpha \times L_{\text{val}} + \beta \times L_{\text{aro}} + (1 - \alpha - \beta) \times L_{\text{exp}} \tag{2.1}$$

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 [56], α and β are determined in order to maximize the performance for the target attribute. In this 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. Overall, the main configuration parameters used for model training are reported in Table 2.3.

Table 2.3 Main configuration parameters for facial emotion recognition model training.

Parameter	**Value**
Model architecture	SENet (pre-trained on VGGFace2 and adapted to predict dimensional emotions)
Trainable parameters	32,343,802
Input size	224 × 224 pixels (RGB color)
Output	Valence (floating point in $[-1, 1]$), arousal (floating point in $[-1, 1]$), categorical expression (integer in [0, 7])
Dataset size	AffectNet—291,650 images (training: 273,269; validation: 14,382; test: 3999)
Data augmentation	Random horizontal flip and rotation $\pm$ 20°
Optimizer	Adam
Learning rate	Adaptive (starting from 10^{-4} and halved after 5 epochs of no gain in the validation loss)
Batch size	32
Steps per epoch	250
Maximum number of epochs	90
Loss function	Multi-task learning weighted loss (combining mean squared error losses related to valence and arousal, and a cross-entropy loss related to the categorical expression, as defined in Eq. (2.1))
Early stopping	After 10 epochs of no gain in the validation loss
Hardware	Local server with NVIDIA GeForce RTX 3080 GPU, Intel i9-10900X CPU, 64 GB RAM

Model Evaluation

As commonly used and formulated in [31], the performance evaluation on the test set is expressed in terms of the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2} \tag{2.2}$$

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

$$\text{CCC} = \frac{2\rho\sigma_{\hat{\theta}}\sigma_{\theta}}{(\mu_{\hat{\theta}} - \mu_{\theta})^2 + \sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2}; \rho = \frac{\text{COV}\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{2.3}$$

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 [31]. 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.

2.3.3 Data Collection Protocol

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 [57]. Therefore, an emotion-eliciting video is set up, using images and sounds from IAPS (International Affective Picture System [58]) and IADS-2 (International Affective Digitized Sounds-2 [59]), 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 [60]. 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 the current 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 2.7. This protocol is inspired by that in [61].

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 [48] is used to set up the experiment. It allows 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.

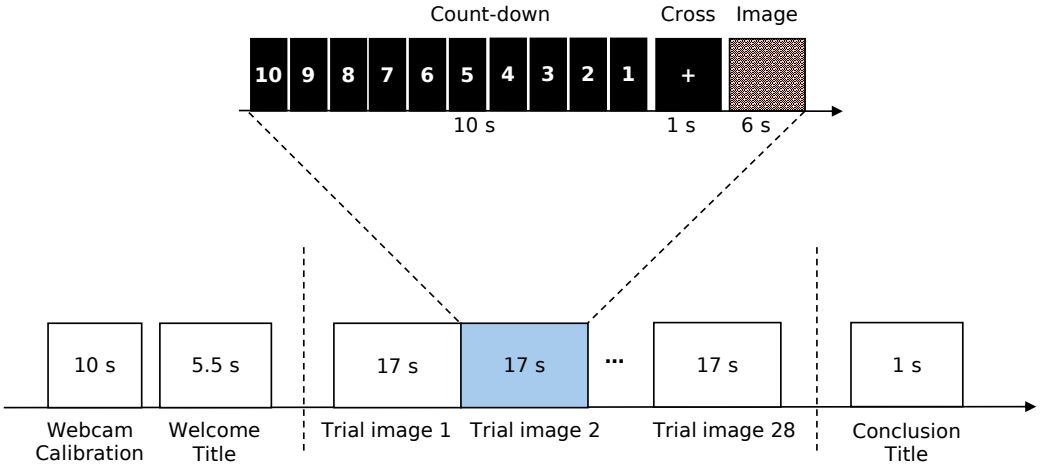


Fig. 2.7 Structure of the emotion elicitation video protocol, inspired by [61] and adapted from [1].

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 the whole emotion elicitation setup, which is inspired by the setup of Prajapati et al. [62], is shown in Figure 2.8. 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 [63].

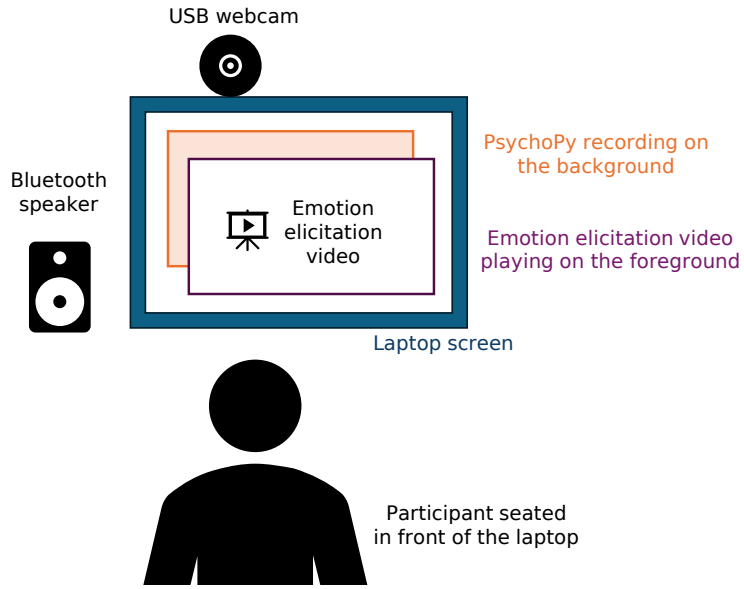


Fig. 2.8 Illustration of the emotion elicitation setup, inspired by [62] and adapted from [1].

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 181). 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.

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 ≥ 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 first phase of 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 2.4.

Table 2.4 Demographics and relevant clinical data of the participants in the first phase of the experiment.

	Cognitively Impaired	**Healthy Controls**
Number of subjects	32	28
Age (mean $\pm$ 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 $\pm$ standard deviation)	12.7 ± 5.0	15.6 ± 4.8
MMSE score (mean $\pm$ standard deviation)	23.9 ± 5.3	29.2 ± 1.2
MoCA score (mean $\pm$ standard deviation)	18.7 ± 5.1	25.4 ± 2.2
Severity of cognitive impairment	23 MCI, 9 overt dementia	No cognitive impairment

In the second phase of the experiment, additional video recordings were collected to expand the database. In total, data from 64 participants were obtained, including 28 HC, 26 subjects diagnosed with MCI (13: due to AD; 13: other types), and 10 diagnosed with overt dementia (4: AD; 6: other types). Table 2.5 provides a summary of the demographic characteristics and key clinical information of the participants in the second phase of the experiment.

Table 2.5 Demographics and relevant clinical data of the participants in the second phase of the experiment.

	MCI	**Overt dementia**	**Healthy controls**
Number of subjects	26	10	28
Age (mean ± standard deviation)	68.2 ± 9.3	72.9 ± 3.8	58.8 ± 6.9
Sex (number of females, %)	10 (38.5%)	6 (60.0%)	14 (50.0%)
Ethnicity	Caucasian	Caucasian	Caucasian
Years of education (mean ± standard deviation)	13.7 ± 4.6	10.4 ± 5.4	15.6 ± 4.8
MMSE score (mean ± standard deviation)	25.8 ± 3.6	18.8 ± 5.5	29.2 ± 1.2
MoCA score (mean ± standard deviation)	20.0 ± 4.4	14.0 ± 3.6	25.4 ± 2.2
Differential CI diagnosis	13: due to AD; 13: other types	4: AD; 6: other types	No cognitive impairment

Diagnoses of AD were performed according to the NIA-AA (National Institute on Aging and the Alzheimer's Association) AT(N) criteria, incorporating biomarkers for amyloid (A), tau (T), and neurodegeneration (N) [11, 64, 65]. Non-AD forms of CI included different etiologies such as FTD, DLB, VD, mixed or not specified forms; individuals with subjective CI were also included. Diagnoses of FTD, DLB, VD and other neurocognitive disorders were made based on established diagnostic criteria specific to each condition [66–69].

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.

2.3.4 Classification of Cognitively Impaired and Healthy Control Subjects

As illustrated in Figure 2.3, the CNN models trained for valence and arousal prediction are used to obtain the subjects' emotional states for all videos in the collected dataset. In the first phase of the experiment, these emotional data are used to train an ML model to classify CI vs. HC subjects.

In the second phase of the experiment, the same workflow is adapted to perform different CI detection tasks. In other words, it still consists of two subsequent parts: (i) obtaining the evolution of the emotions from the collected videos, in terms of valence and arousal; (ii) using the emotions data to train a machine learning (ML) model for the selected CI detection task. A schematic overview of the developed system, including the updates made in the second phase of the experiment, is provided in Figure 2.9. The first steps do not change; only the ML tasks in the last step are different, since not only CI vs. HC differentiation is performed, but also other classification tasks (including MCI vs. HC, and AD vs. other types of CI).

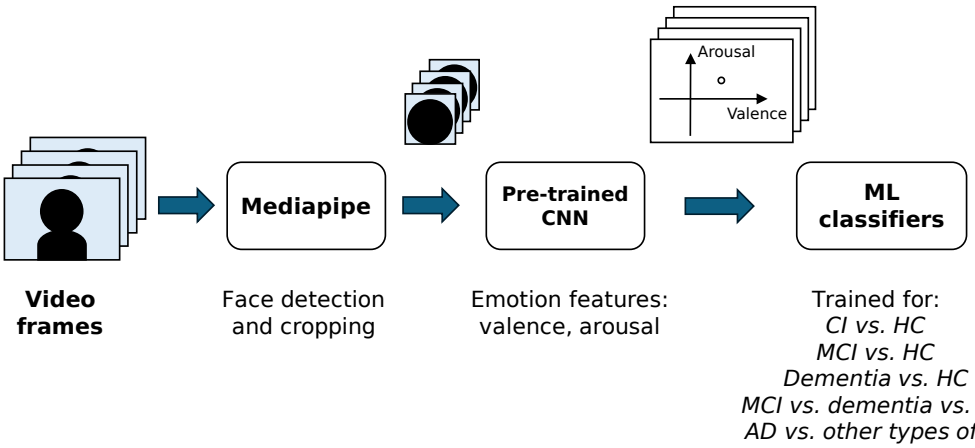


Fig. 2.9 Schematic overview of the proposed system architecture in the second phase of the experiment.

Emotional State Detection

All frames are extracted from the recorded videos ($\sim$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 [70], an open-source framework employing ML to detect face and pose landmarks in real time from video data. The Google Mediapipe Holistic (GMH) solution provides a total of 543 landmarks per frame (33 pose landmarks, 21 hand landmarks for each hand, and 468 face landmarks) [71]. In the presented workflow, GMH is used exclusively for automatic face detection and cropping. Specifically, the minimum confidence value for landmark detection and tracking is set to 0.8. The face bounding box is computed from the facial landmarks by taking the minimum and maximum landmark coordinates along the horizontal and vertical axes.

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.

Classification Tasks

In the first phase of the experiment, the focus is on CI vs. HC classification on the collected video recordings from 60 participants, including 32 CI subjects and 28 HC (described in Table 2.4).

In the second phase, with the same workflow outlined in Figure 2.9, five different experiments are performed on the enlarged version of the dataset (described in Table 2.5):

1. *CI vs. HC*. In this experiment, all CI subjects are grouped together and compared to the HC group through a binary classification task. This enables the validation of the generalization capability of the proposed algorithm when tested on the expanded dataset, compared to that in the first phase of the experiment. The used dataset includes 64 subjects in total: 36 CI (26 MCI + 10 overt dementia) and 28 HC.

2. *MCI vs. HC*. For this experiment, only the subjects with a clinical diagnosis of MCI are selected, among the CI group. The objective is to investigate whether any differences from the HC group would be detected during the earlier stages of the disease. The dataset includes 54 subjects: 26 MCI and 28 HC. A binary classification task is applied to distinguish between these two classes.

3. *Dementia vs. HC*. In contrast to the previous experiment, this analysis includes only patients with overt dementia, with the aim of identifying the differences from the HC group appearing during the later stages of the disease. The dataset includes 38 subjects: 10 overt dementia and 28 HC. A binary classification task is carried out to distinguish between these two classes.

4. *MCI vs. dementia vs. HC*. In this experiment, the three different classes of subjects are compared, according to the level of severity of the disease. The dataset includes 64 subjects in total: 26 MCI, 10 with overt dementia, and 28 HC. The analysis moves from a binary to a multiclass classification task among the three classes. It should be noted that the dataset is imbalanced across classes, with the overt dementia group including fewer subjects compared to the other two.

5. *AD vs. other types of CI*. The aim of this last experiment is to investigate any differences in facial emotion responses among individuals with different types of CI. Specifically, patients diagnosed with AD are grouped together and compared to the broader group of individuals with other forms of CI. This approach is motivated by the fact that AD is the most common cause of dementia, and a differential diagnosis distinguishing AD from other etiologies is of critical clinical importance. The dataset includes 36 subjects: 26 MCI (13: due to AD; 13: other types), and 10 subjects with overt dementia (4: AD, 6: other types). Two classes are considered: AD (17 subjects), and other types of CI (19 subjects). A binary classification task is performed to distinguish between these two classes.

Machine Learning Model Selection and Evaluation

For the binary classification tasks, different ML algorithms are tested, using the implementation provided by scikit-learn. The algorithms deemed suitable for our setup (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}).

For the multiclass classification task (MCI vs. dementia vs. HC), the same ML models are implemented similarly, with a few adjustments to accommodate the multiclass setting. As the scikit-learn KNN estimator supports multiclass problems, no modifications are required. For the LR estimator, the "liblinear" solver is changed to "lbfgs", to handle multinomial loss. In the SVM estimator, multiclass classification is managed using a one-vs-one strategy, implemented following the same procedure used for the binary classification tasks.

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.

2.4 Results

This Section presents the results achieved in this study. Firstly, it details the performance evaluation of the proposed model for facial emotion recognition (Section 2.4.1). Secondly, it introduces the cognitive impairment detection results obtained in the first phase of the experiment, where the focus is on the CI vs. HC discrimination (Section 2.4.2). Lastly, it shows the results obtained in the second phase of the experiment, where the enlarged dataset is used, and the ML tasks include also the discrimination of different stages and etiologies of CI (Section 2.4.3).

2.4.1 Facial Emotion Recognition

The performance of the selected CNN models on our test set from AffectNet for different values of α and β is shown in Table 2.6. The results obtained by Mollahosseini et al. [31] 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 Eq. (2.1)). 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 0.433 for valence prediction and 0.364 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 (0.557 vs. 0.541), with a comparable result on arousal prediction (0.445 vs. 0.450). Therefore, our best CNN models demonstrate equivalent performance to the AffectNet benchmark overall. 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.

Table 2.6 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 [31] is also reported. The best results are emphasized in bold.

	Valence		**Arousal**	
	RMSE	**CCC**	**RMSE**	**CCC**
AffectNet benchmark [31]	**0.394**	**0.541**	**0.402**	**0.450**
CNN, $\alpha = 0.4$, $\beta = 0.6$	0.471	0.501	0.391	0.368
CNN, $\alpha = 0.5$, $\beta = 0.5$	0.454	0.511	0.376	0.421
CNN, $\alpha = 0.6$, $\beta = 0.4$	0.466	0.495	0.379	0.410
CNN, $\alpha = 0.4$, $\beta = 0.5$	0.437	0.529	0.379	0.407
CNN, $\alpha = 0.5$, $\beta = 0.4$	0.457	0.514	0.390	0.376
CNN, $\alpha = 0.3$, $\beta = 0.5$	0.442	0.538	0.374	0.416
CNN, $\alpha = 0.4$, $\beta = 0.4$	0.446	0.529	0.383	0.394
CNN, $\alpha = 0.5$, $\beta = 0.3$	0.455	0.524	0.378	0.395
CNN, $\alpha = 0.3$, $\beta = 0.4$	0.436	0.539	**0.364**	**0.445**
CNN, $\alpha = 0.4$, $\beta = 0.3$	**0.433**	**0.557**	0.369	0.449

2.4.2 Cognitive Impairment Detection

Table 2.7 summarizes the NCV results provided by the ML models tested for CI vs. HC classification in the first phase of the experiment, with 60 participants (refer to Table 2.4). 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 2.7 CI vs. HC classification results. The accuracy and F1 scores achieved by the tested algorithms are reported as (mean $\pm$ standard deviation), together with the corresponding optimal parameters.

	Optimal Parameter Combination	**Accuracy**	**F1 Score**
KNN	5 neighbors, Manhattan distance	$\mathbf{0.767 \pm 0.062}$	$\mathbf{0.754 \pm 0.077}$
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

2.4.3 Discrimination of Different Forms and Stages of Cognitive Impairment

The results obtained in the different classification experiments involving CI and HC subjects in the second phase of the experiment (refer to the dataset in Table 2.5) are shown in Table 2.8. In summary, when classifying all CI subjects vs. HC, the best-performing model is a KNN with a 73.6% accuracy, and an F1 score of 72.2%. Instead, when considering different stages of CI separately, another KNN model reaches the highest accuracy of 76.0% and an F1 score of 74.5% when classifying MCI versus HC subjects. On the other hand, an SVM reaches the best accuracy of 73.6% in distinguishing dementia from HC subjects. However, in this experiment, the F1 scores for all the tested models are lower, reflecting a tendency to misclassify dementia subjects, likely due to class imbalance and limited sample size. Lastly,

for the multiclass problem (MCI, dementia, and HC), a cross-validation accuracy of 64.1% is reached by a KNN model. F1 scores in this setting remain relatively low.

Table 2.8 Cross-validation performance for different classification experiments involving CI and HC subjects (mean ± standard deviation).

Experiment	Model	Parameters	Accuracy	F1 score
CI vs. HC	KNN	3 neighbors, Manhattan distance	**0.736 ± 0.102**	**0.722 ± 0.111**
	LR	L2 penalty, tolerance = 0.0001, C = 0.001	0.623 ± 0.139	0.620 ± 0.141
	SVM	linear kernel, tolerance = 0.001, C = 0.01	0.624 ± 0.092	0.612 ± 0.092
MCI vs. HC	KNN	3 neighbors, Manhattan distance	**0.760 ± 0.041**	**0.745 ± 0.048**
	LR	L2 penalty, tolerance = 0.0001, C = 0.001	0.684 ± 0.114	0.674 ± 0.110
	SVM	linear kernel, tolerance = 0.001, C = 0.001	0.667 ± 0.069	0.664 ± 0.068
Dementia vs. HC	KNN	3 neighbors, Euclidean distance	0.732 ± 0.097	0.487 ± 0.156
	LR	L2 penalty, tolerance = 0.0001, C = 0.1	0.654 ± 0.145	0.492 ± 0.174
	SVM	linear kernel, tolerance = 0.001, C = 0.0001	**0.736 ± 0.018**	0.424 ± 0.006
MCI vs. dementia vs. HC	KNN	5 neighbors, Manhattan distance	**0.641 ± 0.103**	0.463 ± 0.076
	LR	L2 penalty, tolerance = 0.0001, C = 0.01	0.591 ± 0.104	0.427 ± 0.109
	SVM	linear kernel, tolerance = 0.001, C = 0.1	0.578 ± 0.077	0.413 ± 0.051

To complement the quantitative analyses, a scatter plot of valence and arousal values averaged across each video is generated to provide an intuitive visualization

of the emotional information contained in the feature vectors (Figure 2.10). The obtained valence–arousal distribution is consistent with the experimental design: given the deliberate avoidance of highly intense stimuli to ensure the safety and comfort of participants, the elicited responses naturally cluster around moderate values, without reaching extreme levels. This constrained affective space contributes to the observed overlap among groups, although some group-level differences can still be detected. On average, both the MCI and dementia groups display slightly lower valence values compared to HC, whereas arousal values remain comparable across the three groups. Overall, this visual overlap underscores the intrinsic difficulty of the classification task and highlights the necessity of ML approaches to capture more subtle and multidimensional patterns in the data.

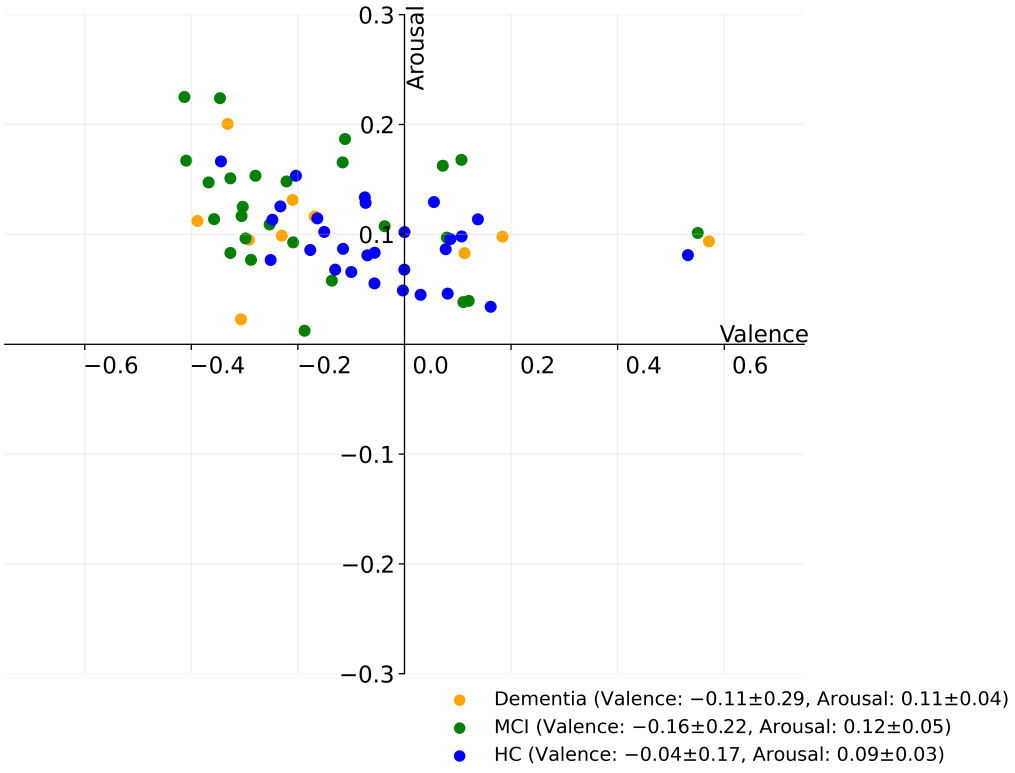


Fig. 2.10 Scatter plot of valence and arousal values averaged across all frames of each video. The legend also reports the group-level summary statistics (mean ± standard deviation of valence and arousal) for HC, MCI, and dementia groups (adapted from [2]).

Table 2.9 reports the results of the classification between AD and other types of CI (experiment 5). As observable, a KNN model achieves the best accuracy of 75.4%, and an F1 score of 74.9%.

Table 2.9 Cross-validation performance for the classification of AD vs. other types of CI (mean ± standard deviation).

Experiment	Model	Parameters	Accuracy	F1 score
AD vs. other types of CI	KNN	5 neighbors, Chebyshev distance	**0.754 ± 0.128**	**0.749 ± 0.130**
	LR	L2 penalty, tolerance = 0.0001, C = 0.0001	0.586 ± 0.171	0.571 ± 0.174
	SVM	linear kernel, tolerance = 0.001, C = 0.01	0.643 ± 0.090	0.602 ± 0.088

2.5 Discussion

For what concerns the emotion recognition step, 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 [31] and is sufficient to be used within our system for cognitive impairment detection.

It is worth recalling that the objective of this research is 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 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 [43, 72] and visual transformers [73, 44]. 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.

As for cognitive impairment detection, the achieved results in the first phase of the experiment (see Table 2.7) demonstrate that the proposed system 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. Due to the limited size of the dataset, the reported accuracy should be interpreted as a promising preliminary result within a proof-of-concept context. 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.

Considering the results obtained in the second phase of this study, they confirm that the employed models are capable of distinguishing CI and HC subjects using solely facial emotion data collected with our protocol (Table 2.8). With respect to the results of the first phase (reporting a 76.7% accuracy for the classification of CI vs. HC subjects), the second phase, based on an enlarged dataset, achieves a comparable accuracy of 73.6%, suggesting that the method maintains good performance despite increased heterogeneity within the CI population. Nevertheless, further data collection remains fundamental to more comprehensively assess the generalizability of these findings.

The results obtained for CI detection support the potential of emotional features extracted from recorded video data as an effective tool to aid CI screening. This further highlights the value of exploring the emotional dimension in the context of early diagnosis, an area currently underexplored in standard neuropsychological assessments such as the MMSE and MoCA, yet shown to be informative in previous studies [16, 21, 22].

When analyzing the classification performance across different stages of cognitive decline (specifically, MCI vs. HC, and dementia vs. HC) the accuracy is comparable to or even higher than that observed in the CI vs. HC classification, with values of 76.0% and 73.6%, respectively. The observed difference between the MCI vs. HC (76.0%) and dementia vs. HC (73.6%) accuracies should be interpreted cautiously, as a more reliable comparison would require similar sample sizes for both groups. Nevertheless, these preliminary results demonstrate satisfactory accuracy when focusing exclusively on MCI patients, without those with overt dementia.

Notably, the F1 scores are consistently lower for the dementia vs. HC classification, suggesting a greater tendency toward misclassification in the dementia group. This may be due to class imbalance and the greater heterogeneity of facial expressions in more advanced stages of cognitive decline. However, the enrollment of overt dementia patients is difficult, especially in advanced stages. Overall, these findings suggest that the proposed method holds promise for supporting the early diagnosis of cognitive impairment, as it performs effectively in subjects at an earlier stage of CI.

Regarding the multiclass classification task (MCI vs. dementia vs. HC), the accuracy decreases compared to the binary classification tasks (MCI vs. HC, and dementia vs. HC), yet it remains relatively high (64.1%) considering the increased complexity of the problem. This outcome is not surprising, as distinguishing CI individuals from HC is generally easier than the more fine-grained differentiation between different stages within the dementia continuum. In addition, the F1 scores obtained in the multiclass setting are lower overall, indicating greater classification uncertainty and class imbalance effects when attempting to separate three groups simultaneously. Interestingly, this limitation was also observed in studies applying ML to neuroimaging data for the diagnosis and prognosis of CI and dementia. In fact, Pellegrini et al. [74] reported that although ML methods achieve an acceptable accuracy in distinguishing overt AD from HC, their performance drops when tackling the differentiation of MCI from AD, MCI from HC, or the prediction of MCI conversion to AD.

One of the most notable findings of this study regards the discrimination of AD vs. other types of CI, reaching a cross-validation accuracy of 75.4% (Table 2.9). This result is particularly promising, suggesting that a differential diagnosis of AD might be feasible through a non-invasive approach, by exploiting facial emotion analysis.

A comprehensive performance comparison with related studies is currently challenging, due to the differences in the video datasets used, the corresponding data collection protocols, and the varying definitions of the ML classification tasks. A distinguishing feature of the present work, compared to others such as [16, 21], lies in the use of an emotion-eliciting protocol based on standardized and extensively validated stimuli [58, 59], enhancing its ease of adoption and generalizability. Moreover, the present study adopts a dimensional model of emotions, providing a more comprehensive characterization of affective states compared to the categorical approach used by Fei et al. [16], who also focused on CI detection based solely on

facial emotions. Most importantly, one of the key strengths of the present study is the availability of well-characterized ground truth classifications for CI subjects. Indeed, unlike other studies where type of CI of the subjects was often assumed and not properly confirmed [23, 15], in this presented study the diagnosis process is based on a comprehensive evaluation, as explained in Section 2.3.3, and AD diagnoses are supported by biomarkers from cerebrospinal fluid.

While the experimental results are promising, this study presents with some limitations. First, the collected dataset has a limited sample size (64 participants), including exclusively individuals of Caucasian ethnicity and recruited from a single clinical center within a specific setting. This is because 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. To further validate the generalization capability of the presented results, future work will focus on expanding the dataset, possibly including diverse ethnicities, and in a multicenter perspective. This will allow to assess the robustness of the proposed method across different populations and clinical contexts, in view of a possible future clinical deployment.

Second, when considering three distinct groups of participants in the second phase of the experiment (HC, MCI, dementia), the dataset is imbalanced across classes, with the dementia group including less subjects than the MCI and the HC. This imbalance results from the recruitment process, in which patients were enrolled based on eligibility criteria during outpatients visits to the Center for Alzheimer's Disease and Related Dementias, without prior knowledge of their clinical diagnosis. Additional experiments with larger datasets and the adoption of more advanced techniques to address class imbalance will be performed in future studies. Future developments will also investigate suitable data augmentation strategies to address the limitations related to sample size and variability.

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 [22]. 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. Moreover, transformer-based end-to-end models such as MC-ViViT [14] (introduced in Section 2.2.1) have recently shown promising results in distinguishing video segments of MCI subjects from those of

cognitively healthy individuals. The ViViT [75] backbone, in particular, processes short video clips to extract rich spatio-temporal features, allowing the model to integrate both appearance and temporal dynamics in its predictions. Exploring these architectures represents a promising direction for future research in CI detection, as they may capture subtle temporal patterns in facial expressivity that frame-level CNN approaches cannot fully exploit.

Another direction for future improvements involves artefact-aware video processing techniques, as facial artefacts, such as involuntary movements, may affect emotion recognition. In this study, data collection was carried out in a controlled environment, which helped to limit extreme occlusions or noise; moreover, the processing 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.

Lastly, a promising direction for future work would be the integration of facial expression analysis with physiological signals, collected for example through non-invasive wearable devices such as wristbands. Smartwatch devices like the Empatica EmbracePlus [76] align well with the proposed data collection framework, as they can be easily added to the designed protocol. Previous studies have demonstrated that combining facial and physiological features (such as galvanic skin response, skin temperature, and heart rate) enhances emotion recognition performance [77, 78]. Consequently, adopting a multimodal approach could further improve the robustness and generalizability of the proposed system for emotion recognition and cognitive impairment detection.

2.6 Conclusion

In this research, 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.

In a second phase of the experiment, the feasibility of using the proposed system to discriminate between different stages and etiologies of dementia is investigated. The dataset size is increased, collecting video recordings of up to 64 participants exposed to the same emotion elicitation protocol. Facial emotion features in terms of valence and arousal are extracted and used to train machine learning models on multiple classification tasks, including distinguishing individuals with MCI and overt dementia from healthy controls, and differentiating Alzheimer's disease from other types of cognitive impairment. The system achieves a cross-validation accuracy of 76.0% for MCI vs. HC, 73.6% for dementia vs. HC, and 64.1% in the three-class classification (MCI vs. dementia vs. HC). Among cognitively impaired individuals, a 75.4% accuracy is reached in distinguishing AD from other etiologies. These results demonstrate the potential of AI-driven facial emotion analysis as a non-invasive tool for early detection of cognitive impairment, and for supporting differential diagnosis of AD in clinical settings.

Chapter 3

Automatic Pain Assessment in the Pediatric Emergency Department

This work has been carried out in collaboration with S.C. Pediatria e Neonatologia - A.O. Ordine Mauriziano di Torino, S.C. Pediatria d'Urgenza - Ospedale Infantile Regina Margherita, and Scuola di Specializzazione in Pediatria - Università degli Studi di Torino. Part of the work described in this Chapter has been published in a conference paper: Bergamasco et al. (2023), *Measurement of Acute Pain in the Pediatric Emergency Department Through Automatic Detection of Behavioral Parameters: A Pilot Study* [79].

3.1 Introduction

Acute pain is a frequent and feared symptom in childhood and is reported in up to 78% of admissions to the Pediatric Emergency Department (PED) [80]. Pain should be adequately considered, measured, and treated whenever it is reported by children or their caregivers, regardless of age, clinical situation, and social role [81]. Acute and repetitive pain experienced at early stages of life can lead to persistent structural and functional changes of the nociceptive system, helping to determine the final architecture of pain system [82]. Many studies have indeed confirmed that untreated pain produces both short and long-term physical and psychological negative effects: healing times are lengthened, complications increase, and long-term sequelae may

develop [83, 84]. Accurate assessment and measurement of pain are the cornerstones of pain management and are essential to provide timely adequate analgesic strategy.

The measurement of pain in children under the age of 3 years, who cannot provide effective self-assessment, requires standardized, validated tools appropriate to their developmental level, the context, and their prior pain experiences. This can be mostly challenging in the PED setting, in particular on the very first evaluation in triage, where time is limited, anxiety is high, and children and their caregivers are unfamiliar with healthcare professionals and environment [85].

Available validated objective scales are based on the evaluation of both physiological and behavioral parameters. However, literature data suggest that the actual use of algometric scales in the PED is limited. Major critical issues are related to environmental factors specific to triage, heterogeneity of scales used, and training deficiencies [86–88]. Among objective pain scales, the Face, Legs, Activity, Cry, and Consolability (FLACC) is based on the detection of behavioral parameters and has been validated for children less than 3 years also in the emergency setting [89]. Due to such reasons, the assessment of pain through objective scales in younger children could be improved by the development of automated machine-based systems aiming to monitor different pain indicators and providing a consistent, minimally biased evaluation of pain.

In this context, the present work aims to address some of the identified gaps, exploring the feasibility of implementing an automated approach to pediatric pain assessment in the PED. The main contributions of this pilot study are summarized as follows:

- An automated computerized tool for pain evaluation in children aged less than 3 years is developed, using only video recordings acquired from an RGB camera without the aid of sensors on the skin.
- A dataset and ad hoc registration setting are created to demonstrate the feasibility of the usage of such automatic system for pain assessment for children in this specific age in the PED environment.
- The scores of the behavioral parameters assigned by the automatic system are compared with those assigned by a healthcare operator to the items Face (F), Legs (L) and Activity (A) of the FLACC pain scale, analyzing the potentiality and limitations of the proposed approach.

3.2 Background

In the past years, there has been an increasing interest in the use of technology for understanding human behavioral responses to pain based on the analysis of facial expressions, and of body or head movements, which are the most important indicators in patients with verbal communication inability [90–92]. Other studies have shown that machine-based systems can be used to detect and analyze physiological changes associated with pain, such as changes in skin color, increase in heart rate [93, 94], and changes in the cerebral hemodynamic of specific brain's regions [95]. Several machine-based approaches have been introduced to analyze infants' body movements for the purpose of diagnosing a specific disease [96, 97]. The development of a machine-based multimodal pain assessment tool that dynamically measures pain in infants has been also proposed, based on the analysis of different behavioral and physiological indicators [98]. Anyway, to our knowledge, no clinical trials have yet been conducted, and most of face detection algorithms are designed and trained for adult faces [98].

Our group has already developed and demonstrated a proof-of-concept computerized tool for the evaluation of pain in newborns, based on the analysis of facial expressions in video recordings [99]. Pain scores obtained from automated analysis have been compared to those assigned by trained healthcare professionals according to three objective neonatal pain scales: the Neonatal Facial Coding System (NFCS), the Premature Infant Pain Profile (PIPP), and the Douleur Aiguë du Nouveau-né (DAN), showing that manual pain evaluation is challenging and often results in a large variability across scores between different operators, making automated assessment desirable [99].

3.3 Materials and Methods

3.3.1 Data Collection

For this pilot study, healthy children aged 3-36 months were enrolled among the ones admitted to the PED of Regina Margherita tertiary teaching Children's Hospital (Turin, Italy) between April and September 2022 with acute pain as main or accompanying symptom. The study excluded children with chronic disorders; those for

whom face and limbs were not fully visible because of dressing, medications, or any medical device; and those admitted with high triage priority code. Pain was measured in all children by the same healthcare professional using the Italian validated version of the FLACC scale [89, 100] (see Table 3.1), along with the assessment and recording of heart rate and oxygen saturation for 60 seconds using the pulse oximeter available in the PED. At the same time, a 60 second video of children's full figure was recorded with an RGB camera with a resolution of 1920x1080 pixels and a frame rate of 30 fps. The camera device was placed 100 cm far in front of the subject, in the same light conditions. Written informed consent was obtained from all parents of the involved children. The study protocol was approved by the hospital's Local Ethic Committee.

Table 3.1 Face, Legs, Activity, Cry and Consolability (FLACC) scale (adapted from [89]).

Categories	**Scoring 0**	**Scoring 1**	**Scoring 2**
Face	No particular expression	Occasional grimace/frown, withdrawn or disinterested	Frequent/constant quivering chin, clenched jaw
Legs	Normal position or relaxed	Uneasy, restless, tense	Kicking or legs drawn up
Activity	Lying quietly, normal position, moves easily	Squirming, shifting back and forth, tense	Arched, rigid or jerking
Cry	No cry	Moans or whimpers, occasional complaint	Crying steadily, screams or sobs, frequent complaints
Consolability	Content and relaxed	Reassured by occasional touching, hugging or being talked to, distractible	Difficult to console or comfort

Overall, 14 Caucasian children were recruited (7 males). The median age of the subjects was 19 months (range: 3–36 months), and the age distribution was the following: 1 subject was 3-6 months; 5 subjects were 7-12 months; 4 subjects were 13-24 months; 4 subjects were 25-36 months. A total of 22 minutes of recording were acquired, with a mean length of 1.16 minutes per child. The acquisition of recordings for a sufficient time was possible for all children, with variable duration of recording fragments suitable for automatic analysis, due to the movements of the child.

3.3.2 Pain Scale Implementation

Since the FLACC pain scale has been validated for children less than 3 years in the emergency setting, it is selected as the reference pain scale for the output of the presented automated pain evaluation system. However, in the proposed implementation of the FLACC, cry (C) and consolability (C) categories are excluded, since the audio signal acquired in the PED environment turns out to be too noisy to be properly processed and analyzed. Therefore, for this research study a partial version of the FLACC (that we will call pFLACC) is used, including only face (F), legs (L) and activity (A). The pFLACC score related to an observation period is calculated as defined in Eq. (3.1):

$$pFLACC_{score} = F_{score} + L_{score} + A_{score} \tag{3.1}$$

where F_{score}, L_{score} and A_{score} indicate the scores computed for the single categories (Face, Legs and Activity, respectively) in the considered observation period. Each single-category score ranges between 0 and 2, thus resulting in a total $pFLACC_{score}$ ranging between 0 and 6.

The computation of these single-category scores is based on the analysis of facial expressions and body movements. Some face and body parameters are identified, deriving from the way the FLACC score is traditionally computed, as described in Table 3.1. In particular, the identified parameters are 7: mouth opening, brow bulging, eye squeezing, legs outstretching, pedaling, body movement, and arms flailing.

For each parameter, an algorithm is developed to quantify it using numerical values, as will be better explained in Section 3.3.3. Moreover, a suitable threshold is defined for each parameter to discriminate the range of values of the parameter representing the "normal" condition, and the range of values representing the "non-normal" condition, related to the pain experience. Threshold values are defined empirically and are calibrated based on the available dataset of videos.

All the 7 parameters are continuously measured by the automatic system along all the observation period, at the end of which the $pFLACC_{score}$ is computed. In the case of the videos of the collected dataset, the observation period is 60 seconds, corresponding to the duration of the videos. The values of each parameter are compared with the threshold established for that parameter for the duration of the whole observation period; two time thresholds are defined empirically (1/3 and 2/3 of the observation time) to check for how long the parameters are in "normal" and "non-normal" range, and therefore assign value 0, 1 or 2 to the pFLACC items. The detailed description of the computation of single-category scores (F, L, A) is reported in the following paragraphs.

Face Score

The computation of F_{score} involves mouth opening, brow bulging and eye squeezing parameters. Such parameters are computed and compared to the relative thresholds along an observation period, as previously defined. A score ranging from 0 to 2 is assigned to F_{score}, according to the methodology illustrated in Table 3.2.

Table 3.2 Computation of the face score (F_{score}).

Status of involved parameters	**Interpretation**	**Assigned F_{score}**
Mouth opening, brow bulging and eye squeezing exceeding the corresponding thresholds for less than 1/3 of the observation period	Neutral expression	0
Mouth opening, brow bulging or eye squeezing exceeding the corresponding thresholds for more than 1/3 and less than 2/3 of the observation period	Occasional frown	1
Mouth opening, brow bulging or eye squeezing exceeding the corresponding thresholds for more than 2/3 of the observation period	Frequent frown	2

Legs Score

The computation of L_{score} involves legs outstretching and pedaling parameters. Similarly to F_{score}, such parameters are computed and compared to the relative thresholds along an observation period. A score ranging from 0 to 2 is assigned to L_{score}, according to the methodology illustrated in Table 3.3.

Table 3.3 Computation of the legs score (L_{score}).

Status of involved parameters	**Interpretation**	**Assigned L_{score}**
Legs outstretching and pedaling exceeding the corresponding thresholds for less than 1/3 of the observation period	Relaxed	0
Legs outstretching or pedaling exceeding the corresponding thresholds for more than 1/3 and less than 2/3 of the observation period	Restless	1
Legs outstretching or pedaling exceeding the corresponding thresholds for more than 2/3 of the observation period	Kicking or outstretching legs	2

Activity Score

The computation of A_{score} involves body movement and arms flailing parameters. Also in this case, similarly to F_{score} and L_{score}, the involved parameters are computed and compared to the relative thresholds along an observation period, resulting in a score ranging from 0 to 2. The adopted methodology is illustrated in Table 3.4.

Table 3.4 Computation of the activity score (A_{score}).

Status of involved parameters	**Interpretation**	**Assigned** A_{score}
Body movement and arms flailing exceeding the corresponding thresholds for less than 1/3 of the observation period	Normal position	0
Body movement or arms flailing exceeding the corresponding thresholds for more than 1/3 and less than 2/3 of the observation period	Squirming	1
Body movement or arms flailing exceeding the corresponding thresholds for more than 2/3 of the observation period	Jerking	2

3.3.3 Face and Body Parameters Computation

As mentioned in Section 3.3.2, there are 7 parameters involved in the computation of the $pFLACC_{score}$: mouth opening, brow bulging, eye squeezing, legs outstretching, pedaling, body movement, and arms flailing. These parameters are calculated starting from the child's face and body landmarks, that are detected and tracked in the video recording along the observation period.

To do so, in this work Google Mediapipe Holistic (GMH) is used, an open-source framework using machine learning techniques to detect and track face and pose landmarks in real time from the video recording of a person [70, 71]. GMH is selected as the landmark-extraction framework because it provides dense facial, body, and hand landmarks within a unified architecture while maintaining real-time performance on standard hardware. Recently, several pose-estimation frameworks such as YOLO11 Pose Estimation [101] and DETR-Pose [102] have demonstrated excellent

performance on benchmark datasets and are suitable for real-time applications on standard hardware. However, these models typically provide only a limited set of skeletal keypoints and do not include a full set of high-density facial landmarks. For the purposes of this study, a detailed and anatomically meaningful characterization of facial movements is required, and GMH offers a comprehensive and fine-grained landmark representation that is essential for infant facial analysis.

As introduced also in Chapter 2 of this thesis, for each frame of a video, GMH provides 543 landmarks (33 pose landmarks, 21 hand landmarks for each hand, and 468 face landmarks) [71]. It also allows setting a minimum confidence value for the detection and tracking of these points; in the proposed system, these values are set to 0.5, so that only landmarks with confidence higher than 0.5 are kept as valid ones. For a set of landmarks, we will call a frame where those landmarks are valid a "valid frame". The 0.5 confidence threshold for landmark detection and tracking in GMH is chosen because it provides a balanced trade-off between detection reliability and the risk of discarding valid frames. For what concerns the format of coordinates of the landmarks, the coordinates that GMH provides in the image reference frame are selected. Each landmark is identified by a set of 3D coordinates (x, y, z); in this analysis, only x and y coordinates are used, and not the z coordinate, being an estimation of the depth of the point made by GMH. Figure 3.1 shows some example of frames of a video in the collected dataset, where landmarks provided by GMH are visualized.

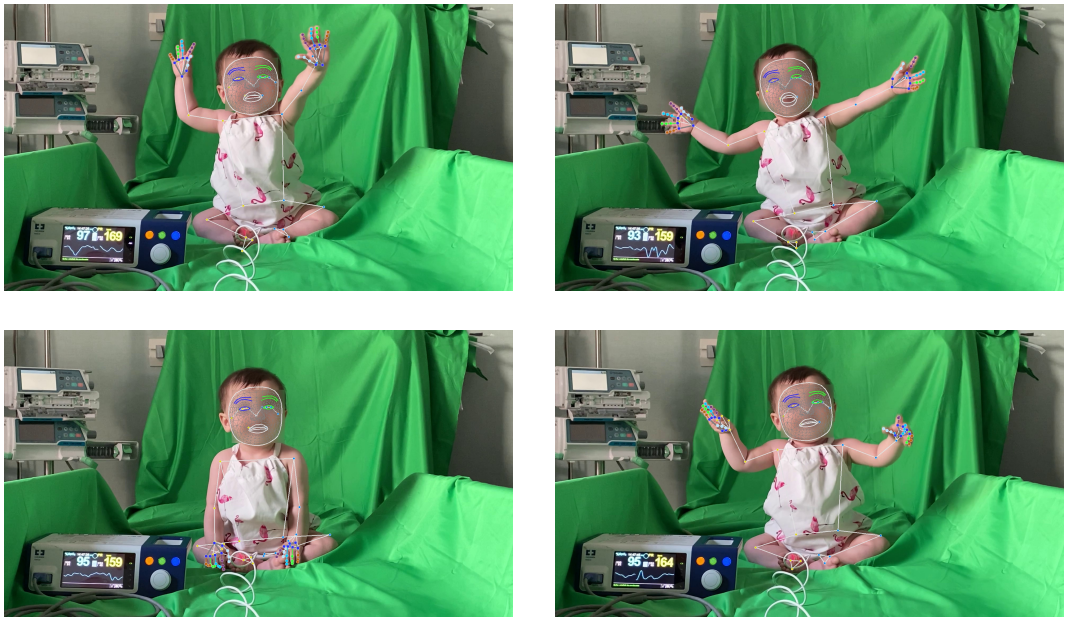


Fig. 3.1 Example frames of a video in the collected dataset, where landmarks provided by GMH are visualized. The images have been postprocessed to anonymize the child's face.

Among all the provided landmarks, the ones judged as the most appropriate to quantify the 7 identified parameters are selected, as will be explained in the following paragraphs. In general, all the parameters are computed with a moving window mechanism, with a window length of 1 second. Basically, each video recording is read frame by frame in a simulated real-time fashion. At each new frame that is being analyzed, the current window is shifted by one frame. Some parameters are calculated for each single valid frame (mouth opening, brow bulging, eye squeezing, legs outstretching), and are therefore averaged over the whole window to get a single value for the window; the other parameters (pedaling, body movement, arms flailing), instead, are calculated only once for the whole window, by considering the variation of intermediate features along the window.

Mouth Opening

For each frame, the height of the mouth is computed as the height of the minimum bounding rectangle enclosing all the mouth landmarks. Similarly, the height of the whole face is computed as the height of the rectangle enclosing all the face oval landmarks. The mouth opening value for the considered frame is given by the ratio between the mouth height and the face height, converted to a percentage. The mouth

opening value for a window is the average of the mouth opening values of the valid frames in the window.

Brow Bulging

For each frame, the distance between eyebrows medial borders is calculated and the width of the whole face is retrieved as the width of the rectangle enclosing all the face oval landmarks. The brow bulging value for the considered frame is given by the ratio between the distance between eyebrows medial borders and the face width, converted to a percentage. The brow bulging value for a window is the average of the brow bulging values of the valid frames in the window.

Eye Squeezing

For each frame, and for both the right and left part of the face, the distance between mid-eyebrow and mid lower eyelid is calculated. Then, the mean between these two distances is retrieved. The face height is computed in the same way as for the mouth opening parameter. The eye squeezing value for the considered frame is given by the ratio between the calculated mean distance and the face height, converted to a percentage. The eye squeezing value for a window is the average of the eye squeezing values of the valid frames in the window.

Legs Outstretching

For each frame, the angle formed by each leg is computed from the positions of the landmarks representing the extremities of the leg and the knee. The legs outstretching value for the considered frame is given by the maximum between the angles formed by each of the two legs. The legs outstretching value for a window is the average of the legs outstretching values of the valid frames in the window.

Pedaling

For each frame, the distance between hip and ankle for each leg is computed from the positions of the corresponding landmarks. In a window, the standard deviation

of these distances is calculated for each leg separately. The pedaling value for a window is given by the maximum of these two standard deviation values.

Body Movement

For each frame, the center of the torso is calculated as the center of the quadrilateral represented by the four landmarks at the extremities of the torso. In a window, the standard deviation of the positions assumed by the center of the torso is computed. The body movement value for a window is given by such standard deviation value.

Arms Flailing

For each frame, the position of the landmarks representing the left wrist and right wrist is considered. In a window, the standard deviation of the positions assumed by each of these two landmarks is computed. The arms flailing value for a window is given by the maximum between the two standard deviation values.

3.4 Results

Section 3.3 illustrated how it is possible to implement a partial version of the FLACC algometric scale to perform automatic pain assessment. In this Section, the comparison between the pain scores given by the healthcare professional on the collected dataset and the corresponding pain scores given by the automatic system is presented. These results will be then analyzed and discussed in the following Section, to investigate the feasibility of usage of the system with children aged less than 3 years in the PED environment.

Table 3.5 reports the scores for each item (F, L, A) and the total pFLACC scores obtained for all the collected video recordings, as assigned by the healthcare professional and the automatic system. For a quantitative comparison, cosine similarity values are also reported in Table 3.6.

Table 3.5 Scores assigned by the healthcare professional and the automatic system to all the video recordings collected in our dataset.

	Healthcare professional				**Automatic system**			
ID	**F**	**L**	**A**	**pFLACC**	$\mathbf{F}_{\text{score}}$	$\mathbf{L}_{\text{score}}$	$\mathbf{A}_{\text{score}}$	$\mathbf{pFLACC}_{\text{score}}$
1	0	0	0	0	0	0	1	1
2	0	0	1	1	0	0	1	1
3	2	1	1	4	1	0	0	1
4	2	0	1	3	2	1	0	3
5	2	0	1	3	2	0	0	2
6	0	0	0	0	0	0	1	1
7	1	0	0	1	0	0	0	0
8	0	0	0	0	0	0	1	1
9	2	1	1	4	1	0	0	1
10	2	1	1	4	0	0	0	0
11	1	0	1	2	0	0	1	1
12	0	1	0	1	0	1	0	1
13	0	0	0	0	1	0	0	1
14	0	0	0	0	1	0	1	2
15	1	0	0	1	0	1	0	1
16	0	0	0	0	0	0	1	1
17	0	0	0	0	0	0	1	1
18	0	0	0	0	0	0	1	1
19	0	0	0	0	0	0	1	1

Table 3.6 Cosine similarity results for the comparison of the scores assigned by the healthcare professional and the automatic system to all the video recordings collected in our dataset.

F	L	A	pFLACC
0.72	0.29	0.24	0.58

3.5 Discussion

As shown in the previous work from our group [99], where a proof-of-concept tool for the automatic evaluation of pain in newborns was developed, while comparing scores obtained from an automatic system and scores assigned by healthcare professionals, we should not aim to merely maximize the agreement between them, but we should rather analyze the causes of the differences between the two scores. This is because of the subjectivity and inter-operator variability that are present in pain evaluation made by healthcare professionals, which are motivating us in the investigation of automatic pain assessment. Therefore, what we can analyze in our comparison are the differences between the values measured by the machine and the healthcare professional for each item, and the reasons behind them. Indeed, for the same observed limb movements and facial expressions, the analysis of what allows the human operator to distinguish pain from restlessness and anxiety is the basis for developing an efficient computer system.

For what concerns the evaluation performed by the healthcare professional, from the comparison of the scores and the feedback given by the healthcare professional, a sort of bias has emerged among analyzed items: facial expression seems to be the most influent to allow healthcare professionals to distinguish restlessness from real pain, followed by activity and lastly by legs movements. In fact, the values assigned to L (legs) and A (activity) quantify movement in general, but do not define the mood in detail. It is therefore difficult to distinguish when this occurs because of pain or due to other reasons.

With respect to the automatic system, instead, the three items (F, L, A) are calculated in a completely independent way, and they have the same weight in the summation that is performed to get the pFLACC score, as described in Eq. (3.1). If there is a high value for the face item, this does not influence the computation of

the legs and activity items. These observations suggest that future investigations could be carried out on the possibility of assigning different weights to the face, legs and activity items in Eq. (3.1), so that facial expression has a higher impact on the pFLACC score with respect to legs and body movements.

On the technical side, the major limitation that can be observed for the automatic system is that the robustness in the landmarks tracking performed by GMH may not be sufficient when movements occur. In fact, when the child turns to the side or moves the head, the landmarks tracking experiences a degraded performance or is even lost. This may lead to get parameters values that exceed the corresponding thresholds, and, consequently, F_{score}, L_{score} and A_{score} assume values greater than 0 even when the child is quiet. This limitation is also related to the choice of the confidence threshold used in GMH. While the adopted value of 0.5 avoids excessive loss of valid frames, it also prioritizes continuity of tracking over stricter detection precision. Higher thresholds (e.g., 0.7) could potentially reduce the number of spurious or unstable landmark estimates; however, this would need to be balanced against the increased likelihood of lost detections during rapid movements or non-frontal head poses, conditions that are common in infants. Exploring this trade-off could be of interest in future developments.

In addition, using adaptive thresholds instead of fixed thresholds for the face and body parameters could make the system more robust to the variation of facial proportions in children of different ages, and help mitigating the effects of the landmark tracking issues.

Comparing this study with the experience of Parodi et al. [99] with newborns, it is worth noting that pain evaluation in video recordings of children aged less than 3 years is more challenging: newborns not only move less, but also have a different range of facial expressions and movements than infants and older children. Such observation could be considered to enhance the software performance and for the future development of the automatic system, focusing on improving the robustness of landmarks tracking in movement condition and the encoding of face parameters.

Of course, the automatic tool tested in this preliminary research is still not suitable to be used in routine practice, as it represents the first exploratory step within a larger research work. Indeed, the performed analysis is limited by the small population, in particular the low number of infants in the collected dataset. Next steps include

the collection of data from other infants, to build a larger dataset that will allow an improved statistical analysis of the results obtained with the proposed system.

Anyway, the achieved results represent the basis to develop a system in which human assessment could be integrated, standardized, and improved thanks to an automatic system. Pain assessment in the PED is an essential part of triage evaluation and is considered as the fifth vital sign [88]. Rapid and standardized assessment is crucial, but environmental and cultural factors may limit optimal performance on most occasions [86–88]. The standard practice to evaluate pain in children who are not able to make self-assessment is based on the observation made by caregivers and healthcare professionals through validated scales that sometimes fail to meet psychometric standards and requires continuous monitoring. Inter-operator variability is undoubtedly a limit of traditional pain measurement with validated scales. In clinical practice, a more objective approach for pain assessment is desirable to correctly recognize and treat pain.

Different approaches for automated recognition of pain expression have been proposed in the last 20 years; however, most of face detection algorithms have been designed and trained for adult faces and are poorly suitable for infants and young children [98, 103–105]. This research proposes an automatic system for objective and contactless pain assessment through the automatic detection of behavioral parameters from video recordings, showing the potentiality that machine-based pain evaluation has also in infants and children aged less than 3 years. To the best of our knowledge, there are no other similar studies in the literature that have approached the development of a system of this kind in the context of the PED.

However, on one side the automatic system could reduce the variability in the use of behavioral pain scale in infants and young children, such as FLACC, and make it faster. On the other side, the expertise of trained triage healthcare professionals will continue to be irreplaceable, as it allows to understand and capture aspects that are not yet appreciable by the machine [99]. In fact, a trained observer can pick up nuances beyond the mere detection of specific movements and expressions, and care-givers can usually discern between face expressions of pain, discomfort, or restlessness of their children. In front of a facial grimace (or a frowning expression), the software calculates movements and expressions in an objective way, while a human operator not only analyzes the same parameters, but is also able to better contextualize them. Therefore, it is essential to properly combine both automatic

and human pain assessment. Understanding those aspects paves the way for further development and improvement of our pain assessment system in the future.

3.6 Conclusion

This pilot study explores the possibility to provide automatic, objective and minimally biased assessment of pain in young children, supporting the evaluation made by healthcare professionals, that remains irreplaceable.

The results of the study suggest that the proposed automated pain assessment system is a promising tool that can potentially aid and improve the evaluation of healthcare professionals even in the emergency setting, providing the basis for further research in this field. Anyway, even when it will be validated for routine practice, automatic detection of behavioral parameters should be never intended as a substitute of the observation provided by healthcare professionals. Instead, it could integrate human evaluation and contribute to make it easier and faster.

The main strength of this study is that, for the first time, it has created the assumptions to collect a dataset for the development of an automatic pain detection system in children in the PED. Moreover, the recording setting has been implemented, adapting the one previously elaborated for newborns [99] to children aged less than 3 years. The collaboration between clinicians and engineers has allowed to create a multidisciplinary project for the development of this system.

Further challenges will be the substantial extension of the original recording dataset and the complementation of video processing with the analysis of audio information, thus implementing the full FLACC pain scale. Moreover, future research will include the improvement of the robustness of landmarks tracking in video recordings of children in movement conditions. Another challenge will be to improve the automatic pain assessment approach so to better encode the nuances that the human operator grasps, and integrate them with the detection of physiological parameters, in order to provide a more objective and standardized evaluation of pain.

Chapter 4

Automatic Pain Assessment in Neonatal Clinical Practice

This work has been carried out in collaboration with S.C. Pediatria e Neonatologia - A.O. Ordine Mauriziano di Torino. Part of the work described in this Chapter has been published in a conference paper: Bergamasco et al. (2024), *Pain Assessment in Neonatal Clinical Practice via Facial Expression Analysis and Deep Learning* [106].

4.1 Introduction

Newborns, even if premature, feel pain and remember pain experiences [107]. Studies have proved that repeated painful exposures during the neonatal period may lead to adverse effects, both in the immediate time (e.g., cardiovascular changes and increased energy expenditure) and in the long term (e.g., increased pain sensitivity and poorer brain development) [108–111]. Accurate and reliable pain assessment using validated tools is crucial for determining the most effective pain management strategies and diminishing the pain-related complications.

Since newborns cannot verbally express their pain experience, several observational scales have been published in literature, which were validated for different newborn populations (term, preterm) and different types of pain (e.g., procedural pain, postoperative pain) [111]. Some of the most commonly used scales are the Neonatal Facial Coding System (NFCS) [90], which evaluates the presence of 9

facial movements; the Premature Infant Pain Profile (PIPP) [112], which combines the assessment of behavioural state, physiological parameters, and facial expressions; and the Douleur Aiguë du Nouveau-né (DAN) [113], which considers facial expressions, limb movements and vocal expressions. However, their use in clinical practice is often limited, or not appropriate for the type of pain and population considered. Moreover, it lacks objectivity as it highly depends on the observer's training and sensitivity, as discussed also in Chapter 3.

For these reasons, newborn pain assessment could greatly benefit from the development of reliable computer-based methods, that would allow a more objective evaluation. Moreover, these automated approaches should be contactless and non-invasive, to avoid complications as skin damages or increased risk of spreading infections.

Some studies have proposed automated approaches to perform pain detection based on the analysis of vocal expressions [114, 115], body movements, facial expressions [116], or multimodal approaches that combine some of these parameters [117, 118]. However, vocal expressions and body movements can hardly be analyzed with pain management measures requiring that newborns are given a pacifier and are wrapped in a blanket. Analyzing only facial expressions, instead, allows contactless pain detection in a way compatible with these pain management measures. In particular, facial parameters as brow bulge, eye squeeze, nasolabial furrow can be exploited, since they are exhibited by the vast majority of term newborns undergoing acute painful procedures, and not when they are subject to other non-painful stimula [90, 116].

Video-based automated systems offer reproducibility and allow to overcome the inter-operator variability in pain assessment [99]. On the other side, they may have limitations in real-time situations, especially when part of the infant's face is occluded or when movements occur [79]. The identification of the most informative regions for neonatal pain diagnosis would allow to assign weights to the visible facial regions and perform pain detection also when a part of the face is occluded [116].

In this context, the present study aims to advance the development of automated, interpretable systems for neonatal pain assessment. The main contributions of this work are summarized as follows:

- A deep learning framework is implemented for neonatal pain assessment based on facial expression analysis.
- The framework is evaluated on a collected dataset of images of newborns undergoing a blood sampling procedure in a real-world clinical environment, where pain management strategies (such as the use of a pacifier) are applied.
- Explainable artificial intelligence (XAI) methods are employed to identify the most informative facial regions contributing to pain detection. This provides insights toward a more objective and efficient pain assessment process, in which human evaluation is supported and complemented by reliable computer-based methods.

4.2 Related Work

There are few datasets publicly available for the development of automated neonatal pain assessment based on facial expression analysis. The iCOPE dataset [91, 119, 120, 92] includes 204 images of 26 newborns, grouped in 2 classes: the *pain* class contains 60 images of newborns experiencing the painful procedure of heel lancing, while the *nonpain* class contains 144 images of newborns either in rest condition or experiencing a nonpain stimulus (transport from one crib to another, air stimulus on the nose, or friction on the heel surface). Because of the similarity of some facial expressions induced by nonpain stimula with pain expressions, this dataset results to be highly challenging, and is still used by most researchers as a benchmark dataset [121].

Lately, a video dataset called iCOPEvid has been presented in [121], aiming to provide information also on the dynamic patterns of facial expressions. iCOPEvid encompasses 234 videos of 49 newborns, from which 20 s segments are extracted and labelled similarly to iCOPE dataset. Table 4.1 shows a list of some other relevant datasets for infant facial pain detection that are used in literature, including both image and video datasets.

Table 4.1 Relevant literature datasets for infant facial pain detection. GA: gestational age, PR: procedural pain, PO: postoperative pain, h: hours, d: days, w: weeks, m: months.

Dataset	Subjects	Age	Ethnicity	Size	Type	Pain labels
iCOPE [91, 119, 120, 92]	26	18-36 h	Caucasian	204 images	PR	pain/ nonpain
iCOPEvid [121]	49	34-70 h	Diversified	234 videos	PR	pain/ nonpain
NPAD [122]	40 (PR+PO)	GA 32-40 w	Diversified	multimodal data	PR, PO	PR: no/ moderate/ severe pain, PO: 5 states
USF-MNPAD-I [123]	58 (PR+PO)	GA 27-41 w	Diversified	multimodal data	PR, PO	PR: no/ moderate/ severe pain, PO: several levels
APN-db [124]	213	GA 26-41 w / 0-26 w	N/A	>200 videos	PR	0-11 pain levels
FENP [125]	106	2 d - 4 w	Chinese	11000 images	PR	calmness/ crying/ moderate pain/ severe pain
DFENP [126]	106	N/A	N/A	1897 videos	PR	calmness/ crying/ moderate pain/ severe pain
YouTube Immunization [127]	142	0-12 m	N/A	142 videos	PR	0-10 FLACC [89] scores
Youtube Blood Test [128]	63 scenarios	N/A	N/A	55 videos	PR	0-4 NFCS [90] scores

Several studies have been conducted in the last 20 years to assess pain using machine and deep learning methods on the iCOPE dataset. Brahnam et al. [91] reach an accuracy of 88.00% in *pain* versus *nonpain* classification using 10-fold cross-validation. However, in this work samples of the same subject are used in both training and test sets. In [119], a classification accuracy of 82.35% is reached by a Support Vector Machine (SVM) with linear kernel tested on a sample not previously seen. Brahnam et al. [120] extend the work in [91] introducing Neural Network Simultaneous Optimization Algorithm (NNSOA), and using Leave-One-Subject-Out (LOSO) cross-validation as an evaluation method, which provides a more realistic performance estimate for clinical settings. NNSOA achieves 90.20% average accuracy in classifying infants' images in *pain* and *nonpain* categories. Celona et al. [129] claim that they have reimplemented the methodology of [120] for comparison purposes, reaching an average accuracy of 78.94%.

Besides machine learning-based methods using handcrafted features, recent research works have also included deep learning-based features, learnt directly from the data and not depending on the subjective design of human experts. Celona et al. [129] investigate the fusion of hand-crafted and deep learning-based features in neonatal facial pain assessment using the iCOPE dataset. Different feature combinations are evaluated with a LOSO cross-validation approach, and the best results are reached with reduced feature vectors extracted from VGG-Face [130] and MBPCNN [131] pre-trained networks (83.78% average classification accuracy). When considering a single feature at time, the pre-trained VGG face achieves the best performance (82.42% average accuracy).

The effectiveness of transfer learning with pre-trained Convolutional Neural Newtorks (CNNs) for feature extraction in neonatal pain expression recognition is confirmed in [132]. This study shows that deep features extracted using VGG-Face architecture lead to better classification performance with respect to other pre-trained CNN architectures. The Neonatal Convolutional Neural Network (N-CNN) is introduced in [133]. This cascaded architecture reaches 91.00% and 84.5% average accuracy on the presented NPAD dataset and iCOPE, respectively, performing binary pain classification task using 3 non-overlapped training and testing sets. Moreover, experiments carried out with N-CNN achieve superior performance with respect to comparative experiments based on the state-of-the-art CNN architecture ResNet [134] and handcrafted features, respectively.

Recently, some studies have included explainable methods in their works on facial pain detection, to improve trust and effectiveness of their machine or deep learning-based solutions. Carlini et al. [135] propose a framework based on a pre-trained VGG-Face model, and fine-tuned on iCOPE and UNIFESP [136] datasets for binary pain classification. The training on both datasets provides a better performance (89% 10-fold cross-validation accuracy) with respect to using only UNIFESP (72%) or only iCOPE (83%) for training. The classification model is also optimized and embedded in a mobile application. Moreover, an explainable method called Integrated Gradients [137] is applied to the test set of both datasets to identify relevant facial regions for pain assessment. The authors conclude that nasolabial groove, open mouth and tongue protrusion may be the most discriminating regions. Sun et al. [138] exploit the Gradient-weighted Class Activation Mapping (Grad-CAM) [139] to select the most important facial movements to be used to train a regression model for infant pain assessment.

Since the time when this study was conducted, transformer-based video architectures have gained considerable attention, achieving state-of-the-art performance across multiple video classification tasks. In the field of pain analysis, ViViT [75] has been applied to binary pain classification [140] and to multimodal transformer frameworks integrating physiological signals and facial expressions [141]. Likewise, VideoMAE [142] has shown promising results in estimating pain expression intensity from video sequences [143]. These recent models offer the ability to capture rich spatio-temporal facial information and therefore represent a compelling direction for future research in neonatal pain assessment.

Nevertheless, the present work is grounded in earlier research that primarily relies on frame-level analysis. This choice is also motivated by several practical considerations, including the limited availability of large-scale annotated video datasets at the time of the study, and the need for models that remain robust under substantial inter-frame variability, occlusions, and movement-related artifacts commonly encountered in real clinical environments. Leveraging convolutional neural networks therefore allows to build upon well-established and widely validated architectures, while enabling consistent comparisons with existing methods in the literature.

In this study, automatic facial pain assessment is performed on a new dataset collected at the Neonatology Department of the Mauriziano Hospital (Turin, Italy), called Mauriziano Pain Assessment In Newborns (M-PAIN) dataset. It reflects the typical characteristics of the real-world environment of the Neonatology Department, as the adoption of pain management strategies (pacifier), lighting conditions varying during the day, and dynamic newborn positions. Moreover, it aims to overcome some limitations of other literature datasets, as a small number of subjects, and inadequate annotations or protocol documentation [116, 124]. The binary pain classification task is addressed, aiming at distinguishing *pain* and *nonpain* conditions. With respect to discrete pain intensity labels, this approach is more robust to the subjectivity of traditional pain assessment, and provides a means for an initial evaluation of the feasibility of our framework. This also allows to include the iCOPE dataset in the performed experiments, enabling possible comparisons with other works in the field. Most importantly, this work contributes to recent efforts to enhance model transparency and interpretability by using XAI methods.

4.3 Materials and Methods

This Section firstly presents the newly introduced M-PAIN dataset, providing all the details about the data collection protocol and the data preparation process (Section 4.3.1). Then, it describes all the steps developed to perform pain classification, including the selected model architecture and training process, and the employed explainable methods (Section 4.3.2).

4.3.1 M-PAIN Dataset

Subjects

This dataset is originally composed of 79 videos of 79 different full term newborns (31 males), aged 36-78 hours, with different ethnicity: 73 Caucasian, 5 African and 1 Asian. The videos were recorded during the procedure of blood sampling by heel prick, in accordance with the Declaration of Helsinki. Informed consent was obtained from participants' parents, and the study protocol was approved by the Ethic Committee of A.O. Ordine Mauriziano di Torino (approval number CS2/504). The dataset is not publicly available due to privacy reasons.

Recording Setup

The videos were recorded at the Neonatology Department of the Mauriziano Hospital in Turin, Italy. An Intel RealSense Depth Camera D435i was installed over the changing table where newborns underwent blood sampling. The video recordings were performed with 1280×720 resolution at a frame rate of 30 fps, and stored as MKV files. Each video has a different time duration as the duration of the blood collection depends on several factors, e.g., the speed of blood outflow and the behavioural state of the child.

Blood Sampling Procedure

Prior to the withdrawal procedure, the infant was placed on the changing table and was wrapped, as a measure of pain containment. A pacifier was also used to further alleviate discomfort [144]. During the recording, the heart rate and oxygen saturation

of the newborn were measured using a Masimo cardiosaturimeter with an electrode that was placed in one foot and held in place with gauze. The cardiosaturimeter display was recorded for future investigations (no physiological data are used in this study). The other foot of the baby was used by the nurse to take the sample by pricking the heel with a disposable needle and squeezing the foot to encourage blood to flow out. Figure 4.1 shows an example image of M-PAIN dataset during the blood sampling procedure.

Fig. 4.1 Example image of M-PAIN dataset during the blood sampling procedure. The image was postprocessed to anonymize the infant's face.

Pain Assessment

Pain assessment was performed by a healthcare professional using the DAN scale [113] (Table 4.2). This scale is suitable both for preterm and term newborns, and is based on three behavioural indicators: facial expressions, limb movements and vocal expressions. The DAN score is an integer number ranging from 0 to 10, given by the sum of three contributions: facial expressions score (0-4), limb movements score (0-3), and vocal expressions score (0-3).

Table 4.2 DAN pain scale (adapted from [113]). P: period of observation.

Facial expression	**Score**
Calm	0
Whines with half-cycle closing and soft opening eyes	1
Determining the intensity of one or several of the following signs: contraction of the eyelids, frown, or enlargement of nasolabial furrow	
– Light, intermittent, with return to calm ($< 1/3$ P)	2
– Moderate (1/3 – 2/3 P)	3
– Very marked, persistent ($> 2/3$ P)	4
Limb movements	
Calm or gentle movements	0
Determining the intensity of one or several of the following: pedaling, spacing of the toes and lower limbs, stiff raised, flailing arms, withdrawal reaction	
– Light, intermittent, with return to calm ($< 1/3$ P)	1
– Moderate (1/3 – 2/3 P)	2
– Very marked, persistent ($> 2/3$ P)	3
Vocal expression of pain	
No complaining	0
Moaning briefly	1
Intermittent cries	2
Long-lasting cry, continuous scream	3
Total	

An observation window was defined within each video, which started with the heel squeeze and was extended either until the completion of the procedure, or until the infant exhibited signs of calmness. Each video was cut so to keep only the chosen observation window. Then, for each video the pain experienced by the newborn was evaluated by the healthcare professional, who assigned a DAN score, specifying the three contributions (facial expressions, limb movements, vocal expressions).

In this study, specifically, the facial expressions score is considered, ranging from 0 to 4. According to [113], it is based on the evaluation of parameters like eye opening, brow bulging, eye squeezing, and nasolabial furrow. Three videos are

excluded, where the infant's face is either out of frame or excessively covered by a blanket. Therefore, the number of usable videos results to be 76 (50 videos with score 0, 2 videos with score 1, 8 videos with score 2, 11 videos with score 3, and 5 videos with score 4). In this way, the dataset results composed of a limited number of videos exhibiting medium or high pain scores. Hence, a predetermined frame extraction strategy is implemented to prevent a significant class imbalance.

Frame Extraction and Labelling

Since the frame rate of all videos is 30 fps, to reduce redundancy between consecutive frames and avoid an excessive number of nearly identical images in the dataset, a temporally down-sampled frame extraction strategy is applied. For videos featuring a pain score equal to 0 or 1, a frame is selected every two seconds of footage (i.e., one frame every 60 frames is picked). Conversely, for videos characterized by a high pain score (2-4), an extraction rate of one frame per second is chosen (equivalent to one frame every 30 frames). This approach allows for an adequate representation of the expressive content, while minimizing redundancy in low-variability conditions (scores 0–1) and preserving richer information in cases of higher facial expressiveness (scores 2–4). For videos with longer observation windows, only the first part of the window is used for frame extraction, to have almost the same number of frames extracted for every infant.

Within an observation window, the infant usually alternates calm and agitation moments. Indeed, pain scales as the DAN scale evaluate the frequency of facial expressions within an observation window. For this reason, even if a video has been assigned a high pain score, not all the frames within the observation window will exhibit pain-related facial expressions. Following these considerations, all the extracted frames have been manually labelled under the supervision of a healthcare professional. Frames from videos with a pain score of 0 or 1 are categorized as belonging to class *nonpain*; on the other hand, frames from videos with scores ranging from 2 to 4 are classified as belonging to class *pain* or class *nonpain* according to the facial expression present in each frame, as evaluated by the healthcare professional. In the end, a total of 5309 frames is obtained, with 3934 labelled as *nonpain* and the remaining ones labelled as *pain*.

4.3.2 Pain Classification

Experiments Performed

In this study, three experiments are performed. In experiment E1, a pain classification model is built using only the M-PAIN dataset. In experiment E2, the same approach is used on the iCOPE dataset, to enable comparisons with similar research works. Finally, in experiment E3 model training is performed on both M-PAIN and iCOPE, to check if the performance on M-PAIN data can be further improved by the combination of different training data. In all cases, the same transfer learning pipeline implemented in Python is adopted, whose details are explained in the following paragraphs.

Model Architecture

Given the limited size of the datasets used in this study, a transfer learning approach is adopted. The VGG-Face model [130] is chosen to be fine-tuned for binary pain classification, since it has proven to reach the best performance in literature studies. This model has been pre-trained on VGG-Face dataset [130] for the face recognition task. It is based upon the VGG-16 architecture [145], a CNN architecture composed of 5 groups of convolutional layers (each group followed by a pooling layer), and a set of 3 dense layers at the end. The pre-trained VGG-Face model weights are obtained from [53]. In our implementation, the 3 dense layers are removed and substituted with a fully connected classifier composed of 3 dense layers: 2 dense layers with ReLU activation (2048 and 1024 neurons, respectively), and a final two-units layer with softmax activation, which generates binary outputs for *pain* or *nonpain* classes. The training process involves updating the parameters of the last group of convolutional layers and our fully connected classifier, while keeping the remaining layers frozen. To avoid overfitting, dropout layers with a 50% rate are inserted after each dense layer. Overall, the proposed model architecture has 10,230,274 trainable parameters and is illustrated in Figure 4.2.

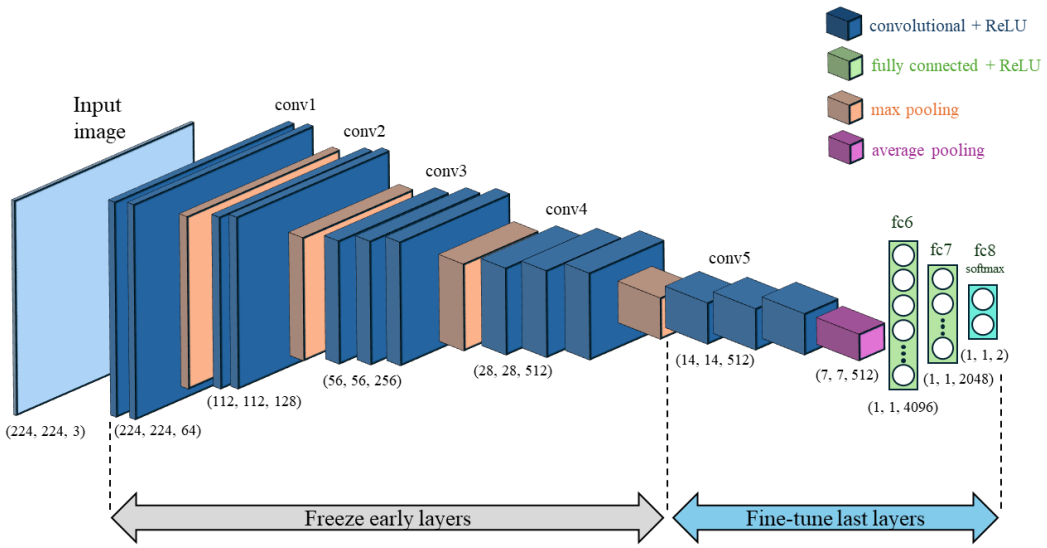


Fig. 4.2 Scheme of the proposed CNN model architecture and transfer learning approach.

Data Preprocessing

The same preprocessing steps are applied in all the experiments regardless of the source dataset of the images, i.e., M-PAIN and iCOPE. Yolov5 with pre-trained weights [146] is employed to detect and crop the subject's face from each frame. This state-of-the-art face detection model was trained on a proprietary dataset [123] containing labelled newborn faces in a clinical setting. Other face detectors trained on adult faces have shown worse performance in recognising newborn faces, due to differences in face composition and expressions if compared to adults [147].

To maintain consistency with the training procedures of the VGG-Face pre-trained model, the same preprocessing method used in the VGG-Face dataset [130] is applied. Thus, the images are converted from RGB to BGR, and each color channel is zero-centered with respect to the mean values of the VGG-Face dataset. All the images are resized to 224×224 pixels, to match the expected dimensions of the pre-trained model. Moreover, data augmentation techniques are applied in the training phases to expand the data and enhance its variability. Data augmentation is limited to include only random horizontal flips, random rotations within a 30-degree range, and brightness adjustments, to avoid introducing unrealistic variations.

Model Training and Performance Evaluation

To prevent overfitting, training is terminated when the validation loss fails to improve for 15 consecutive epochs, and the categorical cross-entropy loss function with L1 regularization penalty is used. Moreover, in the loss function class weights are added to mitigate class imbalance [148], ensuring that the model pays more attention to examples from an under-represented class. In all the experiments the Adam optimizer is used, because it is unaffected by the scaling change induced on the loss by class weights. The batch size is set to 32, which is also the largest feasible value given our available computing resources, and the maximum number of epochs is set to 100. For what concerns the initial learning rate, it is set to $5{\times}10^{-6}$ in experiments involving M-PAIN (E1 and E3). In the E2 experiment involving only iCOPE, a larger value is chosen (10^{-5}) to speed up convergence.

In each of the experiments, to train the pain classification model and estimate its generalization performance, a stratified k-fold cross-validation is employed. This avoids the limitations of a single train-validation-test split, which may not adequately represent the overall population, particularly for small datasets. The stratified k-fold implementation with non-overlapping groups provided by the scikit-learn library [49] is selected. This technique creates folds that preserve the proportion of samples for each class and ensure that no group overlaps between splits. This is important for creating homogeneous training, validation, and test sets and to avoid assigning frames of the same newborn to different sets. At each of the k cross-validation steps, the dataset is partitioned in three parts (training, validation and test sets); a model is trained on the training set, while the test set performance is computed using the model weights obtained on the epoch with the highest validation accuracy. The final cross-validation performance metrics are obtained by averaging the performance of the k trained models.

With M-PAIN data (E1, E3), the dataset size is of 5309 images (74% *nonpain*; 26% *pain*). Therefore, a 5-fold cross-validation is performed, i.e., k=5. With only iCOPE data (E2), instead, 10-fold cross-validation is chosen (k=10) due to the smaller size of the dataset: 204 images (70% *nonpain*; 30% *pain*). Because of class imbalance, the F1-score metric is chosen as an evaluation metric in addition to the accuracy. In this way, it is possible to measure the model's ability to correctly identify images depicting pain while minimizing false positives.

Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) [139] is employed to provide model explainability and transparency. Grad-CAM is a visualization technique used to understand which parts of an image a CNN focuses on when making a prediction. In particular, in this study we follow the implementation in [149]. The Grad-CAM algorithm uses the gradient information flowing into a CNN layer to calculate a weighted sum of the feature maps. Then, it produces a heatmap highlighting the regions in the input image that had the largest influence on the model's prediction on that image. In our implementation, the inputs specified to Grad-CAM at each cross-validation step are the trained classification model, the final convolutional layer (due to its proximity to the classification target), together with the test images and the corresponding model predictions. For each test image, a heatmap is generated, and the results of this step are analyzed in Section 4.4.

4.4 Results

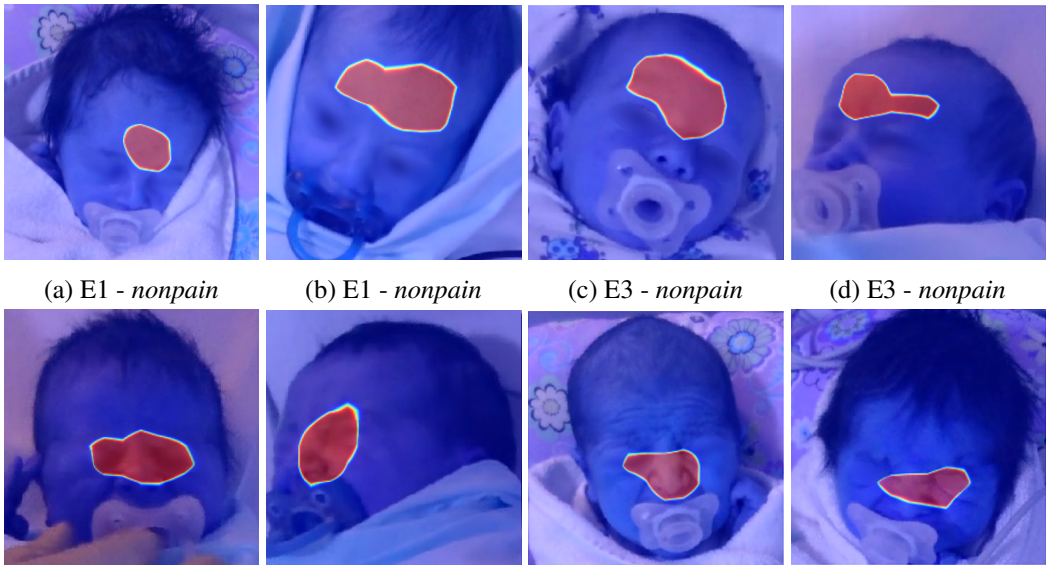
Following the approach described in Section 4.3, three different models for pain detection are trained. Table 4.3 shows the results in terms of mean and standard deviation of the accuracy and F1-score over the cross-validation steps.

A visual examination of the confusion matrices across the five cross-validation folds in E1 indicates that misclassifications do not consistently occur in favour of either the *pain* or the *non-pain* class. Instead, in some folds the model correctly identifies almost all *non-pain* images while showing lower sensitivity to *pain* instances; in others, *pain* images are more correctly classified than *non-pain* ones. These differences likely reflect variations in the distribution, representativeness and difficulty of the samples included in each fold's training, validation and test sets.

Table 4.3 Cross-validation results for *pain* vs. *nonpain* classification, in terms of accuracy and F1-score (mean ± standard deviation).

Experiments performed	**Accuracy**	**F1-score**
E1: training and evaluation on M-PAIN	0.874 ± 0.034	0.754 ± 0.054
E2: training and evaluation on iCOPE	0.838 ± 0.107	0.668 ± 0.293
E3: training on M-PAIN+iCOPE, evaluation on M-PAIN	0.888 ± 0.050	0.796 ± 0.081

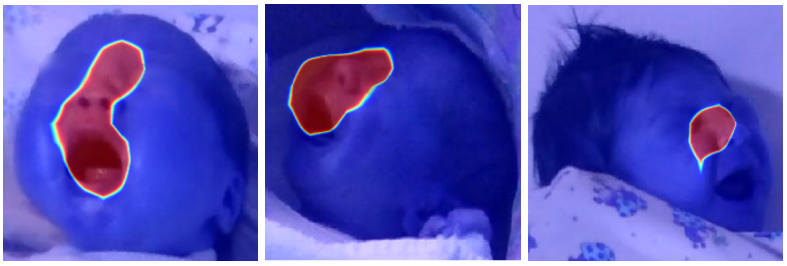
The trained models of E1 and E3 are selected to carry out the interpretability analysis performed through Grad-CAM, in which images from the M-PAIN test set of each cross-validation step are considered. Figure 4.3 shows some examples of the resulting heatmaps. The areas in red are those which most affect the model's prediction. Moreover, Figure 4.4 shows some heatmaps related to frames in E1 where infants have momentarily lost their pacifiers, to enable a comparison of the situation with and without pacifier.



(a) E1 - *nonpain* (b) E1 - *nonpain* (c) E3 - *nonpain* (d) E3 - *nonpain*

(e) E1 - *pain* (f) E1 - *pain* (g) E3 - *pain* (h) E3 - *pain*

Fig. 4.3 Examples of resulting Grad-CAM heatmaps for M-PAIN images. The areas in red are those which most affect the model's prediction. Image labels and experiment numbers are reported below each heatmap. The images have been postprocessed to anonymize infant faces.



(a) E1 - *pain* (b) E1 - *pain* (c) E1 - *pain*

Fig. 4.4 Examples of resulting Grad-CAM heatmaps for M-PAIN images without pacifier. The areas in red are those which most affect the model's prediction. Image labels and experiment numbers are reported below each heatmap. The images have been postprocessed to anonymize infant faces.

4.5 Discussion

The model trained on M-PAIN (E1) using 5-fold cross-validation achieves an average accuracy of 87.4% and an average F1-score of 75.4% in classifying *pain* vs. *nonpain* images. This demonstrates the feasibility of the presented approach for automated facial pain detection, and its potential in supporting human pain assessment for a more efficient evaluation.

Secondly, the model trained on the iCOPE dataset (E2) using 10-fold cross-validation achieves an average accuracy of 83.8% and F1-score of 66.8%. The results of this model present a higher standard deviation, meaning that the performance varies significantly across different cross-validation folds. This can be expected when the dataset size is small, as happens with iCOPE. Moreover, iCOPE is known to be highly challenging due to the presence of *nonpain* images induced by stressful stimuli. Some of these images are indistinguishable from *pain* images with naked eye, even when observed by healthcare professionals.

The results achieved in this study on binary pain classification with iCOPE are comparable to state-of-the-art results in prior studies [119, 129, 133, 135]. However, LOSO method used in other studies [120, 129] may overestimate accuracy on iCOPE, because not all subjects have the same number of images for each category. Hence, the presence of subjects with a limited number of *nonpain* images related to stressful stimula as friction and cry (the most difficult to classify) may lead to an over-optimistic estimate of the model performance. Instead, with stratified k-fold

cross-validation, the test set at each step includes always the same percentage of *pain* and *nonpain* images, leading to a more realistic performance estimation.

Lastly, with a model trained on both M-PAIN and iCOPE (E3), the average accuracy and F1-score on M-PAIN data result to be 88.8% and 79.6%, respectively. Therefore, the cross-dataset training leads to an improvement in both accuracy and F1-score compared to the model of E1, which is trained only on M-PAIN. This suggests that incorporating a wider range of training data can enhance the model performance, particularly in the challenging setting of the Neonatology Department.

With respect to the interpretability analysis shown through the Grad-CAM heatmaps in Figure 4.3, the forehead emerges as the most highlighted facial feature in images correctly classified as *nonpain* (Figs. 4.3a, 4.3b, 4.3c, 4.3d). For images correctly classified as *pain* (Figs. 4.3e, 4.3f, 4.3g, 4.3h), Grad-CAM highlights the upper contour of the nose and nasolabial groove. Notably, the gradients do not emphasize the pacifier or secondary artifacts such as the blanket, focusing primarily on the newborn's face. Therefore, model predictions result to be based on the facial regions most closely associated to the pain experience. This further validates the proposed framework, while revealing interpretability aspects which can provide the basis for an effective integration of automatic pain assessment in clinical practice.

Interestingly, when analyzing frames in E1 correctly classified as *pain* where infants are highly agitated and lose their pacifiers, there are cases in which the mouth area is highlighted in Grad-CAM heatmaps, meaning that the network redirects its focus also to the mouth in absence of a pacifier (Figure 4.4). Even without the pacifier, the nasolabial fold remains highlighted in *pain* images (Figs. 4.4a, 4.4b, 4.4c), in the same way as in *pain* images where the pacifier is present (Figs. 4.3e, 4.3f). On the other hand, the mouth region is often emphasized when not occluded by the pacifier, as it still conveys relevant pain-related information (Figs. 4.4a, 4.4b). Such information cannot be considered for pain detection when the pacifier covers the mouth (Figs. 4.3e, 4.3f).

Building on the insights obtained with Grad-CAM, future extensions could explore more diverse XAI approaches that capture additional aspects of model interpretability within medical imaging. Indeed, while saliency-based ("visual") XAI methods like GradCAM project their explanations directly onto an input image, non-saliency-based ("non-visual") XAI methods provide more diversified explanations [150]. For example, prototype-based networks [151] can learn representative image

parts associated with each class, called "prototypes", and classify new samples by comparing them to these learned prototypes. This mechanism offers transparent, example-based explanations and has shown promise in improving interpretability in medical image classification [152]. Other examples are case-based explanations, enabling "what-if" reasoning through similar samples or altered inputs. In particular, counterfactual explanations modify the input image to produce a different prediction, providing a contrastive example that clarifies the factors driving the model's original decision. This approach has been tested on x-ray image analysis [153].

A natural extension of the present work concerns the transition from binary pain detection to the estimation of multiple pain intensity levels. Moving towards a multiclass setting would be more closely aligned with clinical practice, where graded pain scales as the DAN scale are routinely used for decision-making. Achieving this goal, however, requires a substantial increase in the size of the M-PAIN dataset, while fine-grained pain annotations may be derived from the available DAN pain scores (see Table 4.2).

In parallel, future work could exploit the temporal dimension of pain, taking advantage of the full video sequences rather than treating images as independent frames. Temporal modeling would allow the system to capture the dynamics of facial expressions and behavioral changes over time, which are highly informative for distinguishing pain from other forms of distress or discomfort. As mentioned in Section 4.2, recent transformer-based video architectures, such as ViViT and VideoMAE, offer powerful tools to learn such representations directly from video data and have already shown promising results in pain analysis tasks. Integrating these architectures into future work would represent an important step toward developing real-time pain detection systems capable of continuously monitoring newborns at the bedside and assisting clinicians in their decision-making under realistic neonatal care conditions.

4.6 Conclusion

This study presents a deep learning framework for pain assessment in newborns based on facial expressions. It is developed using image data from M-PAIN, a new dataset of videos recorded in real-world conditions within a Neonatology Department, during neonatal heel prick. In this setup, pain management strategies as the use

of a pacifier are adopted, making the analysis of newborn facial expressions more challenging. The presented approach reaches average accuracy of 87.4% and F1-score of 75.4% using a stratified 5-fold cross-validation on M-PAIN. Moreover, when the training is carried out on both M-PAIN and iCOPE datasets, the performance on M-PAIN data improves to an average accuracy of 88.8% and F1-score of 79.6%. The application of a visual explanation technique based on Grad-CAM shows that the networks' decisions are based on the facial regions most closely associated to the experience of pain, i.e., brow bulge, eye squeeze, nasolabial furrow. These results show that automated pain detection from facial expressions is feasible also in a real-world setup. These promising findings pave the way for the development of an automated system that integrates, standardizes, and improves human pain evaluation. In addition, the proposed approach adds explainability to the neural networks results, enhancing transparency and trust for healthcare professionals. In future research, this work could be extended to a multiclass classification task to infer pain intensity levels, as more data will be collected to further increase the size of M-PAIN. At the same time, future work includes the investigation of the temporal dimension of pain exploiting videos in the M-PAIN dataset, going towards a real-time pain detection system for improved patient monitoring and informed decision-making.

Chapter 5

Large Language Models for Diagnostic Support in the Pediatric Emergency Department

This work has been carried out in collaboration with S.C. Pediatria d'Urgenza - Ospedale Infantile Regina Margherita, Scuola di Specializzazione in Pediatria - Università degli Studi di Torino, and Dipartimento di Scienze Cliniche e Biologiche - Università degli Studi di Torino. Part of the work described in this Chapter has been published in a journal paper: Del Monte et al. (2025), *Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study* [154].

5.1 Introduction

Large Language Models (LLMs) are advanced artificial intelligence (AI) systems that understand and generate natural language [155]. Among the most popular ones, OpenAI's GPT models [156] such as Chat Generative Pre-trained Transformer (ChatGPT) [157], Google's Gemini series [158], and Meta's LLaMA family [159], gained attention in the open-source community. These models are trained on vast amounts of textual data, and their performance improves as the quantity and quality of training data increase [155].

LLMs can be applied in clinical decision support, medical record analysis, patient engagement, and dissemination of health information [160]. AI-based tools can support healthcare professionals by offering diagnostic assistance, thereby increasing accuracy, efficiency and enhancing clinical outcomes [161, 162]. However, sometimes their responses could be inaccurate or misleading, underscoring the need for rigorous validation and oversight in clinical settings [163, 164].

In 2023, Kanjee et al. [165] examined the diagnostic accuracy of ChatGPT-4, showing that AI included the correct diagnosis in differential-diagnosis lists in 64.0% of cases, successfully identifying the main diagnosis in 39.0%. In the same year, Hirosawa et al. [166] evaluated ChatGPT-3 on common clinical scenarios, showing that it included the correct diagnosis in 93.3% of differential-diagnosis lists, though physicians outperformed the model in ranking accuracy. In a follow-up study [167], the same team showed that ChatGPT-4 performed better than ChatGPT-3.5 and comparably to physicians, although the differences were not significant. Recently Hirosawa et al. [168] tested different chatbots on adult cases: ChatGPT-4 achieved the highest accuracy, including correct diagnoses in 86.7% of lists and identifying the main diagnosis in 54.6% of cases.

To our knowledge, the role of LLMs as a diagnostic support tool in the Pediatric Emergency Department (PED) has not been explored yet. In this pilot study, the diagnostic efficacy of some of the most used LLMs is tested on pediatric emergency clinical vignettes, and their performance is compared to a group of physicians. The aim of the study is to evaluate whether LLMs can serve as an effective support to physicians in formulating accurate diagnoses for pediatric emergency clinical cases.

5.2 Background

In order to contextualize the methodological choices of this study, it is useful to briefly outline how modern LLMs are designed. This Section presents an overview that draws on the survey on LLMs presented by Minaee et al. [155], complemented with more recent insights in the field. The intention is not to provide an exhaustive description, but rather to highlight the main architectural components and training principles that are most relevant for understanding the subsequent work.

In particular, firstly LLM building blocks are presented (Section 5.2.1). Afterwards, insights on the usage of these models are reported, such as limitations, prompt engineering, and augmentation techniques (Section 5.2.2). In the end, relevant LLM families are introduced, i.e., the GPT family, the Gemini family, and the LLaMA family (Section 5.2.3).

5.2.1 Building Blocks of LLMs

Model Architecture

Most of the popular architectures for LLMs are based on the Transformer framework, introduced in [169]. Its core is the attention mechanism, allowing a model to dynamically focus on the most relevant parts of the input when processing each element in a sequence. Compared to mechanisms such as convolution and recurrence, attention captures long-range contextual information more effectively.

Current large language models predominantly follow three architectural paradigms: encoder-only, decoder-only (also called auto-regressive models), and encoder–decoder (also called sequence-to-sequence models). Despite their structural differences, these models typically adopt the Transformer as their fundamental computational backbone.

Data Cleaning

The quality of training data plays a decisive role in determining the performance of language models. Preprocessing steps such as filtering and deduplication have been shown to substantially influence the effectiveness and reliability of the resulting models.

Data filtering improves training quality by removing noise, outliers, imbalances, and ambiguities, while also applying text preprocessing to standardize inputs. Techniques range from heuristic rules to classifier-based approaches, with the goal of ensuring cleaner, more balanced, and reliable datasets.

Deduplication eliminates redundant data instances that can bias training, inflate certain patterns, or cause overfitting. This step enhances diversity and generalization,

and is often implemented through document-level similarity measures to detect duplicates.

Tokenization and Positional Encoding

Tokenization is the process of segmenting raw text into smaller units called tokens, which serve as the basic elements for language model training and inference. Tokens may correspond to words, subwords, or characters, depending on the chosen scheme. By converting continuous text into a structured sequence of discrete symbols, tokenization enables models to process, represent, and learn from natural language in a computationally tractable way.

To complement tokenization, positional encoding mechanisms provide sequence order information that Transformers cannot inherently capture. Early approaches used absolute embeddings (e.g., sinusoidal functions), while later methods introduced relative encodings and rotary embeddings to better generalize across lengths and capture pairwise dependencies. More recently, positional bias techniques have enabled extrapolation to longer sequences, improving efficiency and scalability in large LLMs.

Model Pre-training

Pre-training is the initial and most fundamental stage in the LLM training pipeline, as it equips models with broad language understanding capabilities, before task-specific fine-tuning. In this phase, LLMs are trained on vast amounts of unlabeled text, allowing them to learn statistical regularities, syntactic structures, and semantic relationships in a self-supervised manner. The objective is to acquire general-purpose language representations that can later be transferred to a wide variety of downstream tasks.

Several strategies have been explored for pre-training, including tasks such as next sentence prediction. However, the two most widely adopted paradigms are autoregressive language modeling, where the model predicts the next token in a sequence, and masked language modeling, where the model reconstructs masked tokens from their surrounding context. More recently, Mixture of Experts (MoE) architectures have been introduced to scale models efficiently by activating only sub-

sets of parameters per input, thus reducing compute requirements while maintaining or improving performance.

Model Fine-tuning and Instruction Tuning

Fine-tuning adapts pre-trained language models to specific tasks using labeled data, enhancing their effectiveness beyond general-purpose pre-training. While some modern LLMs can already perform tasks zero-shot, i.e., without having been explicitly trained on those tasks, fine-tuning remains valuable for improving accuracy, reducing prompt complexity, and exposing models to domain-specific or proprietary data.

Instruction tuning extends this idea by aligning models with human expectations expressed through natural language instructions. Datasets such as Natural Instructions or methods like Self-Instruct provide structured prompts and examples that guide model behavior. Empirical evidence (e.g., the fine-tuned InstructGPT vs the corresponding foundational model GPT-3) shows that instruction-tuned models consistently outperform their base versions, making this approach central to building reliable and user-aligned LLMs.

A summary table containing the details of several popular pre-training and fine-tuning datasets can be found in [170]. Moreover, the work of Minaee et al. [155] presents a detailed section related to popular datasets used for LLMs, both for basic tasks, such as language understanding, and for emergent LLM abilities, such as reasoning and instruction following.

Alignment

Alignment refers to guiding LLMs so that their outputs are consistent with human goals, values, and preferences, reducing harmful or biased behaviors. Beyond instruction tuning, several approaches have been developed. The most established is Reinforcement Learning from Human Feedback (RLHF), where human-labeled preferences are used to train a reward model that shapes the LLM's behavior. A related variant, Reinforcement Learning from AI Feedback (RLAIF), leverages feedback from another aligned model instead of humans.

A newer method called Direct Preference Optimization (DPO) simplifies the pipeline by directly optimizing on preference data without requiring a separate reward

model, improving stability and performance. More recently, Kahneman-Tversky Optimization (KTO) has been proposed, which further reduces data requirements by relying only on knowledge of whether an output is undesirable, making it more practical and scalable for real-world use.

Decoding Strategies

Decoding denotes the procedure by which pre-trained LLMs generate text. In other words, decoding strategies determine how pre-trained LLMs generate text by selecting the next token from the model's probability distribution.

The input sequence is first processed by a tokenizer, which maps each word or subword into a discrete token and then into its corresponding numerical identifier (token ID). These token IDs are fed into the model, which predicts the probability distribution of the next possible token in the sequence. The model outputs logits, i.e., unnormalized scores over the vocabulary, which are then normalized into probabilities via a softmax function. This probabilistic formulation allows the model to estimate the likelihood of each candidate token, from which the decoding algorithm selects the next token (or sequence of tokens) to extend the generated text.

Various decoding strategies have been introduced. Simple approaches like greedy search always pick the most likely token, but risk incoherence. Beam search expands multiple candidate sequences in parallel, improving quality at higher computational cost. To enhance diversity and avoid repetitive outputs, probabilistic methods are used: top-k sampling selects randomly from the k most probable tokens, while top-p (nucleus) sampling chooses from the smallest set of tokens whose cumulative probability exceeds a threshold. Temperature scaling further controls randomness, allowing a trade-off between determinism and creativity.

5.2.2 LLM Usage

LLM Limitations

Once trained, LLMs may be applied to diverse tasks, with users typically interacting through basic prompt-based queries. Despite their strong capabilities, LLMs still present several intrinsic limitations. They lack persistent memory and cannot retain

information across prompts, making them unsuitable for tasks requiring long-term context. Their outputs are stochastic and probabilistic, meaning the same input can yield different responses depending on training parameters such as temperature. LLMs also have no access to real-time or external knowledge, relying solely on pre-training data, which restricts factual accuracy. Moreover, their large size entails significant computational costs, limiting accessibility. Finally, they are prone to hallucinations, producing outputs that appear plausible but are factually incorrect, a major challenge for reliable deployment. Therefore, to fully leverage their capabilities and mitigate inherent limitations, LLMs often require augmentation through external methods.

Prompt Engineering

Prompt design and engineering is the practice of guiding LLMs through textual inputs, ranging from simple instructions to complex structured prompts. Generally, prompt should contain either instructions or questions; other additional elements are optional. The goal is not only to elicit appropriate outputs, but also to systematically exploit model capabilities while accounting for their limitations and contextual constraints.

Advanced methods include Chain of Thought (CoT), which makes reasoning steps explicit; Tree of Thought (ToT), which explores multiple reasoning paths in parallel; Self-Consistency, which compares multiple outputs to identify the most reliable response; and Reflection, where the model evaluates and revises its own outputs. Other approaches such as Expert Prompting, Chained prompts, and Rails further enhance reliability by simulating expert reasoning, structuring multi-step processes, or enforcing domain-specific constraints. More recently, Automatic Prompt Engineering (APE) has emerged to automate prompt generation, scoring, and refinement. Taken together, these techniques transform prompting from a simple interface into an iterative engineering discipline, designed to improve accuracy, coherence, and trustworthiness of LLM-generated outputs.

LLM Augmentation with External Knowledge and Tools

One of the main limitations of pre-trained LLMs lies in their lack of access to up-to-date or domain-specific information. Retrieval-Augmented Generation (RAG)

addresses this issue by combining LLMs with external knowledge sources such as search engines or knowledge graphs. In RAG pipelines, the model formulates a query, retrieves relevant information from external databases, and incorporates it into its generation process. This approach not only improves factual accuracy but also enables more adaptive and context-aware responses. Recent advances, such as Forward-looking Active Retrieval Augmented Generation (FLARE), further refine RAG by dynamically integrating prediction and retrieval during text generation, thereby enhancing both accuracy and relevance.

In parallel, LLMs can be extended through the use of external tools, which expand their capabilities beyond text prediction. These tools can range from APIs and calculators to specialized services that enable reasoning and decision-making. For instance, tool integration allows models to perform complex tasks like querying databases, executing computations, or interacting with structured knowledge sources. Methods such as Automatic Multi-step Reasoning and Tool-use (ART) have been proposed to coordinate internal reasoning with external tool usage, supporting tasks that require both inference and factual grounding. Together, RAG and tool integration represent a crucial step towards transforming LLMs from static language generators into adaptive systems capable of interacting with dynamic information environments.

Recent advances have led to the emergence of LLM-based agents, which extend these models into autonomous systems capable of interacting with users and environments, making decisions, and executing tasks. Building on external knowledge integration and tool use, LLM Agents take augmentation a step further by not only accessing and leveraging tools, but also making decisions based on context and user interaction, thus enabling more autonomous and adaptive behavior.

5.2.3 LLM Families

The GPT Family

The GPT family, developed by OpenAI, comprises a series of decoder-only Transformer-based language models that have progressively advanced the state of natural language processing. The first generations, GPT-1 [171] and GPT-2 [172], demonstrated the potential of large-scale pre-training on unlabelled text and were released as open-source models, laying the foundation for subsequent research. GPT-3 [173],

introduced in 2020 with 175 B parameters, represented a major leap in scale and capability, showing emergent abilities such as in-context learning, where the model could solve novel tasks from few-shot demonstrations via text interaction with the model, without additional fine-tuning. Derived from GPT-3, Codex [174] extended these capabilities to code generation, while WebGPT [175] fine-tuned the model to interact with information from the web. To better align outputs with user intent, OpenAI introduced InstructGPT [176], trained via RLHF, which significantly improved truthfulness and reduced harmful outputs.

The public release of ChatGPT [177] in November 2022 marked a turning point, popularizing conversational AI systems built initially on GPT-3.5 (a sibling model to InstructGPT), and later GPT models. These models were fine-tuned to follow natural language instructions and power dialogue-based interactions such as question answering, summarization, and information retrieval.

GPT-4 [156] was launched in March 2023, and further extended these capabilities. Unlike its predecessors, GPT-4 is multimodal, capable of processing both text and images as input to produce text outputs. Moreover, it exhibits human-level performance across a variety of benchmarks, including professional and academic exams. Despite being proprietary, GPT-4 has become a cornerstone in the development of large language models, illustrating the effectiveness of combining large-scale pre-training with alignment techniques such as RLHF to produce models that are both powerful and adaptable.

GPT-4o [178], released in May 2024, introduced "omni-modal" capabilities, natively processing text, image, audio, and video. It delivered GPT-4 performance but faster and at lower cost, with notable gains in multilingual and audio-visual understanding. In parallel, in July 2024 OpenAI also introduced GPT-4o mini [179], a lighter and more cost-efficient variant designed to provide strong performance in everyday tasks while maintaining the multimodal capabilities of GPT-4o. Since GPT-4o was the most capable OpenAI model available at the time, and its functionality was exposed to users through ChatGPT-4o, this version was chosen as the basis for our experimental evaluation, along with the cost-efficient ChatGPT-4o mini.

Other newer models were introduced by OpenAI after the period in which our study was conducted. GPT-4.5 [180], released in February 2025 (codenamed Orion), served as OpenAI's last standalone GPT model before the unified system architecture. It contributed enhancements through scaled unsupervised learning, yielding

improvements in pattern recognition, user alignment, and emotional intelligence. The most recent milestone is GPT-5 [181], launched in August 2025. GPT-5 introduced a unified multi-regime architecture with real-time routing, dynamically directing input to specialized reasoning or fast-response experts depending on task needs. It achieved state-of-the-art performance across diverse domains like coding, math, and medical reasoning, solidifying its position at the cutting edge of large language models.

The Gemini Family

The Gemini family, developed by Google DeepMind, represents a succession of highly capable, multimodal large language models designed as successors to LaMDA [182] and PaLM 2 [183]. Introduced in December 2023, the initial lineup included Gemini Ultra, Gemini Pro, and Gemini Nano, each optimized for different use cases: Ultra for maximum reasoning and multimodal performance; Pro for broad versatility and powering Google's chatbot; and Nano for lightweight, on-device applications [158].

Subsequent iterations advanced the family significantly. Gemini 1.5 [184] introduced models such as Gemini 1.5 Pro and 1.5 Flash, integrating both dense and sparse scaling methods and enabling efficiency in reasoning, planning, multilingual understanding, function calling, and long-context processing. These were the selected models for our experiments, since they were the most powerful Gemini models available at the time of our study.

More recently, Gemini 2.0 (December 2024) introduced the Flash series, including Flash and Flash-Lite, designed for low-latency, cost-efficient performance in multimodal tasks. These models brought expanded agentic capabilities and real-time multimodal support including audio and video comprehension. Gemini 2.5 (March 2025) represents a major advance in the Gemini family, offered in two principal variants. Gemini 2.5 Pro is the most capable model to date, setting new benchmarks in coding, reasoning, and multimodal understanding. It operates as a thinking model, performing structured reasoning before generating outputs, and supports extremely long context windows (up to 1 million tokens) as well as extended video inputs of up to three hours. In parallel, Gemini 2.5 Flash provides strong reasoning capabilities while maintaining high efficiency in terms of latency and computational cost. Taken together, the Gemini 2.X models illustrate the full Pareto frontier of capability versus

cost. Therefore, they enable users to select models that best balance efficiency and capability, while also pushing the boundaries of what is achievable in complex, agentic problem solving [185].

The LLaMA Family

The LLaMA family, developed by Meta, represents a suite of open-source foundation language models. Being open-source means that the model weights are released under a noncommercial license; this has enabled widespread use within the research community [155].

The first generation of LLaMA models [159], introduced in February 2023, ranged from 7B to 65B parameters and was trained on trillions of tokens drawn from publicly available datasets. Architecturally, LLaMA builds upon the GPT-3 Transformer design but incorporates several architectural refinements. These modifications, combined with efficient training, enabled strong performance: the 13B parameter model outperformed the much larger GPT-3 (175B) across multiple benchmarks, establishing LLaMA as a solid baseline for LLM research.

In July 2023, Meta, in collaboration with Microsoft, released the LLaMA-2 models [186], including both foundation models and dialogue-oriented derivatives known as LLaMA-2 Chat. The latter models were specifically fine-tuned for conversational applications and were reported to outperform other open-source LLMs on a wide range of public benchmarks, further solidifying the impact of the LLaMA line in the broader landscape of large language models.

The third generation, LLaMA-3, was introduced in April 2024 and represents a significant leap forward. Available initially in 8B and 70B parameter variants, both pre-trained and instruction-tuned, LLaMA-3 was trained on approximately 15 trillion tokens and includes models fine-tuned with over 10 million human-annotated examples. It demonstrated state-of-the-art results, outperforming peer open-source and proprietary models, including Gemini Pro 1.5 and Claude 3 Sonnet, on a range of benchmarks. Meta also released subsequent expansions such as LLaMA-3.1 (adding a 405B model), along with multilingual and multimodal capabilities, a much larger context window (up to 128K tokens), and enhanced coding and reasoning performance [187]. These advances firmly positioned LLaMA-3 as a leading open-source LLM platform.

Most recently, in April 2025, Meta unveiled the LLaMA-4 family [188], its most advanced set of open-weight models at the time of writing this thesis. LLaMA-4 introduces a MoE architecture combined with native multimodal capabilities, enabling unified processing of text, image, and video inputs. However, this collection of models was not available yet during the period in which our study was conducted.

5.3 Materials and Methods

5.3.1 Study Design

This prospective observational diagnostic study was conducted at the Regina Margherita Children's Hospital PED between March and October 2024. This tertiary care teaching hospital provides care for critically ill patients younger than 18 years. The study was performed according to the international regulatory guidelines and current codes of Good Epidemiological Practice.

Two experienced pediatricians from our group created a dataset of 80 cases with varying clinical complexity, from different pediatric subspecialties (Table 5.1). The cases were extracted from anonymized records of children admitted to the PED between September 2018 and May 2024. Trauma and cases in which the final diagnosis was reached mainly through laboratory or instrumental tests were excluded. Patients and their parents did not provide written or oral informed consent, as all the cases were anonymized before the vignettes were generated and no sensitive data was reported. Since it was not possible to trace the identity of the patients and since this study did not retrospectively influence in any way the clinical management of the cases described, the approval of the Ethics Committee was not necessary.

Table 5.1 Clinical cases divided by pediatric subspecialties. In the central column, the total number of cases in that subspecialty is reported; in parentheses there is the number of cases that belong to more than one medical specialty. In the list of clinical cases (right column), those classified in more than one subspecialty are in italics.

Pediatric subspecialty	**Number of cases**	**List of clinical cases**
Respiratory system	8 (4)	*Bronchiolitis*, Pneumothorax, Foreign body inhalation, *Pneumonia*, Bronchospasm, *Acute laryngitis*, Pneumomediastinum, *Whooping cough*.
Infectious diseases	17 (13)	*Bronchiolitis*, *Thyroglossal duct overinfection*, *Acute otitis media*, Periorbital cellulitis, *Pneumonia*, Group A beta hemolytic Streptococcus' acute pharyngotonsillitis, *Otomastoiditis*, *Pyelonephritis*, *Retropharyngeal abscess*, Malaria, Mononucleosis, *Staphylococcal scalded skin syndrome*, *Osteomyelitis*, *Meningoencephalitis*, *Pertussis*, *Staphylococcal toxic shock*, *Acute laryngitis*.
Orthopedics	7 (2)	Painful pronation of the elbow, Transient synovitis of the hip, Legg-Calvé-Perthes disease, Epiphysiolysis, Griesel's syndrome, *Osteomyelitis*, *Osteosarcoma*.
Otolaryngology	5 (4)	*Thyroglossal duct overinfection*, *Acute otitis media*, Laryngomalacia, *Retropharyngeal abscess*, *Otomastoiditis*.
Gastro-intestinal tract	8 (1)	Appendicitis, Intestinal invagination, Chronic inflammatory bowel disease, Cyclic vomiting, Biliary tract atresia, Hirschsprung's disease, Functional abdominal pain, *Alagille's syndrome*.
Oncology	4 (2)	Osteosarcoma, Leukemia, *CNS neoplasm* (2 cases, different clinical presentation).
Endocrinology	3 (0)	Onset diabetes mellitus type 1, Hypothyroidism, Addison's disease.
Hematology	7 (2)	Immune thrombocytopenia, Post-infectious bone marrow aplasia in patient with spherocytosis, Hemophilia, Post-infectious acute hemolytic anemia, Hemolytic crisis in favism, *Retinal thrombosis in autoimmune disease*, *Hemolytic-uremic syndrome*.
Nephrology	4 (2)	Post-infectious glomerulonephritis, *Pyelonephritis*, Idiopathic nephrotic syndrome, *Hemolytic-uremic syndrome*.
Immuno-rheumatology	7 (4)	Kawasaki syndrome, *Sydenham's chorea*, *Rheumatic disease*, Systemic juvenile idiopathic arthritis, Schoenlein-Henoch purpura, *Ataxia telangiectasia*, *Retinal thrombosis in autoimmune disease*.
Neuropsychiatry	17 (5)	Guillain-Barré syndrome, *Charcot-Marie-Tooth disease*, Conversion disorder, Transverse myelitis, Febrile convulsive seizure, Trigeminal neuralgia, CNS demyelinating disease, Convulsive seizures associated with gastroenteritis, Migraine with aura, Iatrogenic peripheral neuropathy, *Meningoencephalitis*, *CNS neoplasm* (2 cases, different clinical presentation), Peripheral paralysis of the 7th cranial nerve, *Ataxia telangiectasia,* Narcolepsy, *Sydenham's chorea*.
Allergology	3 (0)	Cow's milk protein allergy, FPIES, Anaphylactic shock.
Cardiology	5 (2)	Complete atrioventricular block in rare pathology (KSS), Myocarditis/heart failure, Vaso-vagal syncope, *Rheumatic disease*, *Alagille syndrome*.
Dermatology	4 (3)	*Staphylococcal scalded skin syndrome*, Subgaleal hematoma, *Kwashiorkor*, *Staphylococcal toxic shock*.
Dietetics and Nutrition	2 (1)	Scurvy, *Kwashiorkor*.
Genetics	2 (2)	*Alagille syndrome*, *Charcot-Marie-Tooth disease.*
Toxicology	2 (0)	Acute accidental intoxication (by cannabinoids), Methemoglobinemia due to local anesthetic.
Visual apparatus	1 (1)	*Retinal thrombosis in autoimmune disease.*

Each case was used to generate a clinical vignette written in Italian by the two expert pediatricians. The clinical vignettes were prepared both to be input as a prompt to different LLM-based chatbots and to be evaluated by a group of physicians. In each vignette, all the main details were presented as follows: recent and past medical history, relevant family medical history, physical examination, and vital signs. Laboratory tests were not reported. Figure 5.1 shows an example of a clinical vignette translated in English.

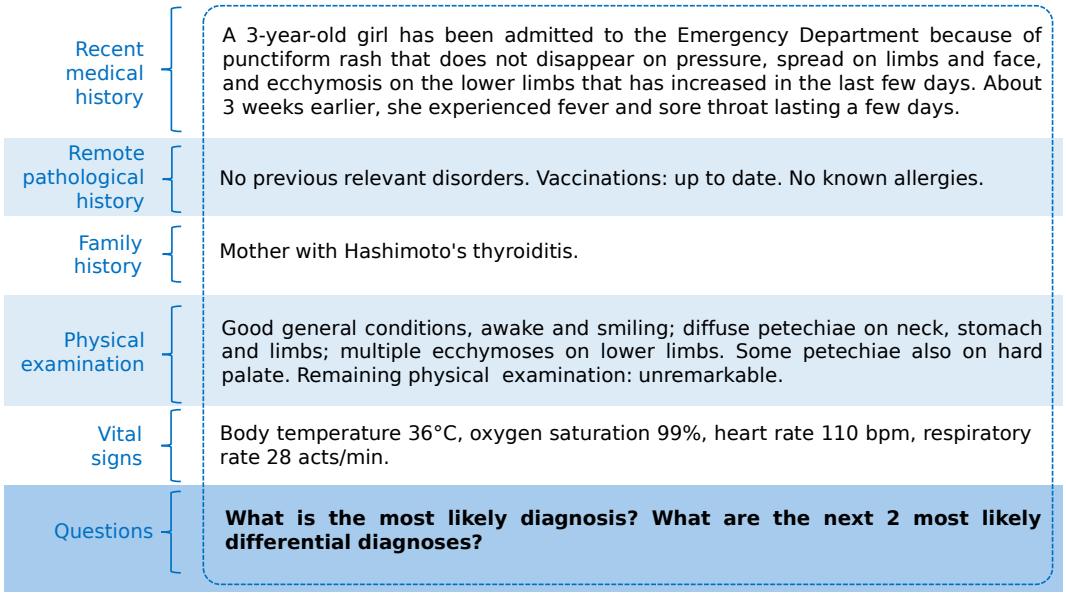


Fig. 5.1 Example of a clinical vignette translated in English, subdivided into its main parts (adapted from [154]).

The vignettes were submitted to a panel of three independent expert pediatricians who validated the cases or recommended a revision. They also independently ranked them according to three difficulty levels (lowly difficult, difficult, and highly difficult), based on solving complexity according only to available clinical data. The final level for each case was determined based on the majority agreement among the experts: 20 (25.00%) highly difficult, 31 (38.75%) difficult, and 29 (36.25%) lowly difficult.

After collecting the answers generated by LLMs and physicians, our two expert pediatricians (the ones that had prepared the vignettes) evaluated all the answers, and statistical analysis was performed. A scheme of the study design is presented in Figure 5.2.

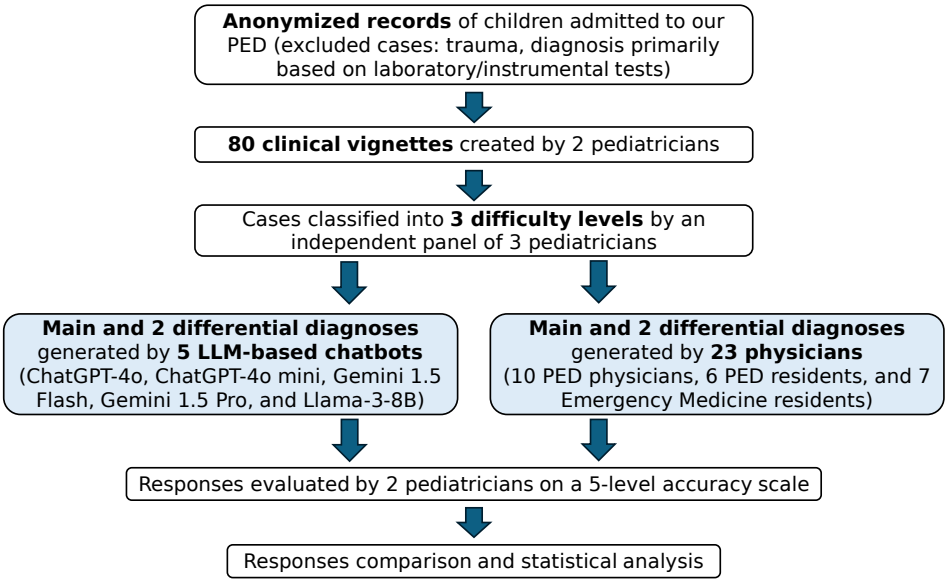


Fig. 5.2 Scheme of the study design.

5.3.2 LLM-based Chatbots' Answers

Four of the highest rated [189, 190] LLMs publicly available during the period in which this study was conducted were selected: ChatGPT-4o [178] and ChatGPT-4o mini [179] (OpenAI); Gemini 1.5 Flash [184] and Gemini 1.5 Pro [184] (Google); and Llama-3-8B [187] Instruct version (Meta), an open-source model satisfying our computational resources constraints. Unlike the other LLMs, which were used through the web interface, Llama-3-8B was deployed in our computing infrastructure and could be used without requiring internet access. The characteristics and access details of the selected LLMs are summarized in Table 5.2 [157, 191–193].

Table 5.2 LLMs selected for the study with their main characteristics.

	ChatGPT-4o [157]	**ChatGPT-4o mini [157]**	**Gemini 1.5 Flash [191]**	**Gemini 1.5 Pro [192]**	**Llama 3-8B [193]**
Provider	OpenAI	OpenAI	Google	Google	Meta
Access date	August, 15	August, 26	August, 21	August, 19-20	August, 27
Open-source	No	No	No	No	Yes
Price (at the time of the study	$20 USD/month	Free after login in	Free after login in	$23.99 USD/month	Free
Training data knowledge cut-off	October, 2023	October, 2023	November, 2023	November, 2023	March, 2023
Release date	May 13, 2024	July 18, 2024	May 14, 2024	February 15, 2024	April 18, 2024
Others	-	-	Internet access	50 questions per day free on Google AI Studio	Deployed on a server with NVIDIA GeForce RTX 3080 GPU (10GB)

A zero-shot approach was implemented, meaning that the chatbots were not provided with any example of the task at hand. Moreover, each vignette was given as a prompt to each chatbot only once in independent chats, to prevent LLMs from applying any learning and inference to subsequent cases. At the end of each vignette, two open-ended questions were asked: "What is the most likely diagnosis? Which are the next two more likely differential diagnoses?".

5.3.3 Physicians' Answers

Twenty-three physicians were selected to evaluate the clinical cases, including 10 PED physicians with at least 5 years of experience, 6 residents attending their last year of residency in Pediatrics at Regina Margherita hospital's PED, and 7 residents attending their last year of residency in Emergency Medicine (EM) at the University of Turin, Italy. These three subgroups were selected to ensure a diverse range of clinical experiences and perspectives. Naturally, the two expert pediatricians in our group who created the dataset and had to evaluate the responses were not included in this recruited group of 23 physicians.

Between July and August 2024, the participants were asked to resolve the 80 vignettes through Google Forms. The use of digital resources, textual assistance or consulting colleagues were forbidden, to ensure that the responses were purely the result of the physician's independent clinical reasoning and experience.

The vignettes were presented in random difficulty order and divided in 4 standardized forms with 20 cases each, in order to minimize the risk of fatigue for participants, thus influencing the quality of responses. As for chatbots, the same questions were asked to physicians for each vignette, i.e., "What is the most likely diagnosis? Which are the next two more likely differential diagnoses?".

5.3.4 Evaluation Method

The answers obtained from LLMs and physicians were independently evaluated by the two expert pediatricians and compared to the final diagnoses established at the time of patients' discharge from the PED or following hospitalization. Each answer was evaluated through a 5-point accuracy scale, in order to avoid penalizing incomplete or imprecise diagnoses that still demonstrated adequate clinical reasoning: 1 (correct main diagnosis); 0.75 (if the correct diagnosis was identified within differential diagnoses); 0.5 (if the main diagnosis was correct, but not precise); 0.25 (if the correct diagnosis was identified within differential diagnoses, but not precise); and 0 (both main and differential diagnoses were incorrect). Examples and explanations for each accuracy category are presented in Table 5.3.

In case of disagreements between the two expert pediatricians, they reached consensus facing each other. Each physician or LLM received a total score by summing the points obtained from all answers, thus 80 was the maximum possible score.

Table 5.3 Examples and explanations for each accuracy category, illustrated using the case of a patient discharged from the PED with a Charcot Marie-Tooth diagnosis. The correct answer, when identified, is written in *italics*.

Accuracy category	**Example answer**	**Explanation**
1	**Principal diagnosis:** *Charcot–Marie–Tooth disease (CMT)* **Differential diagnoses:** Spinal Muscular Atrophy (SMA), Distal Muscular Dystrophy.	The principal diagnosis is identified correctly.
0.75	**Principal diagnosis:** Marfanoid–Hooding Syndrome **Differential diagnoses:** Muscular Dystrophy, *Charcot–Marie–Tooth Syndrome (CMT)*.	The final diagnosis is identified correctly, but as a differential diagnosis.
0.5	**Principal diagnosis:** *Hereditary neuromuscular disease* **Differential diagnoses:** Duchenne Muscular Dystrophy (DMD), other muscular dystrophies.	The correct final diagnosis is identified as the principal diagnosis, but is not precise.
0.25	**Principal diagnosis:** Duchenne Muscular Dystrophy (DMD) **Differential diagnoses:** *Hereditary neuromuscular disease*, Spinal Muscular Atrophy (SMA).	The correct final diagnosis is identified as a differential diagnosis, but is not precise.
0	**Principal diagnosis:** Duchenne Muscular Dystrophy (DMD) **Differential diagnoses:** Becker Muscular Dystrophy (BMD), Spinal Muscular Atrophy (SMA).	All the diagnoses are not correct. The real diagnosis is not identified.

5.3.5 Statistical Analysis

Descriptive statistics are presented using mean ± standard deviation (SD), and median and Interquartile Range (IQR) to report the performance of LLMs and physicians, as appropriate. Bar charts, stacked charts, and dot plots are used to visualize the total scores obtained by the different groups and comparison. For the statistical analysis, a long-format dataset is created. The distribution of accuracy count is checked using histograms and Q-Q plots. Comparisons between physicians and LLMs are made using the Kruskal–Wallis H test. Pairwise comparisons are made using Dunn's procedure with Bonferroni correction for multiple comparisons. Statistical significance is set at $p<0.05$. Analyses are conducted using STATA 18.5.

5.4 Results

Overall, a total of 1,840 responses from the 23 physicians (800 from PED physicians, 480 from PED residents, 560 from EM residents) and 400 responses from the 5 selected chatbots are obtained.

Figure 5.3 illustrates the total scores obtained by all evaluators, grouped by categories (i.e., chatbots and groups of physicians). The scores of each evaluator are also numerically reported in Table 5.4, together with mean values, standard deviations, median values, and IQRs calculated over the three categories of physicians and the chatbot group. The highest and lowest total accuracy scores are obtained respectively by ChatGPT-4o (72.5) and Llama-3-8B (33.75). Gemini 1.5 Flash, ChatGPT-4o mini and Gemini 1.5 Pro score 56.5, 56.75, and 62.75, respectively. PED physicians (60.88 ± 4.83) and PED residents (63.96 ± 2.3) achieve the highest scores, followed by EM residents (44.25 ± 4.64)

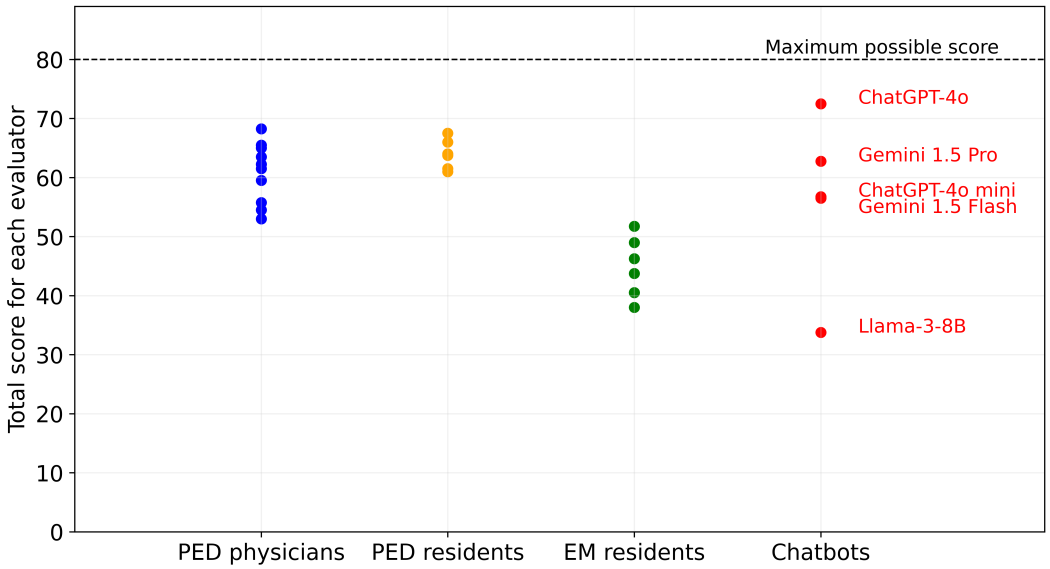


Fig. 5.3 Total scores for each evaluator, grouped by category. PED, pediatric emergency department; EM, emergency medicine (adapted from [154]).

Table 5.4 Accuracy scores of physicians and chatbots (individuals and groups). The maximum achievable score is 80. PED: Pediatric Emergency Department; EM: Emergency Medicine; SD: Standard Deviation; IQR: Inter-Quartile Range.

	Absolute score	**Mean score**	**SD**	**Median score**	**IQR**
PED physician (1)	53	–	–	–	–
PED physician (2)	54.5	–	–	–	–
PED physician (3)	55.75	–	–	–	–
PED physician (4)	59.5	–	–	–	–
PED physician (5)	61.5	–	–	–	–
PED physician (6)	62.25	–	–	–	–
PED physician (7)	63.5	–	–	–	–
PED physician (8)	65	–	–	–	–
PED physician (9)	65.5	–	–	–	–
PED physician (10)	68.25	–	–	–	–
PED physicians (total)	–	60.88	4.83	61.88	56–65
PED resident (1)	61	–	–	–	–
PED resident (2)	61.5	–	–	–	–
PED resident (3)	63.75	–	–	–	–
PED resident (4)	64	–	–	–	–
PED resident (5)	66	–	–	–	–
PED resident (6)	67.5	–	–	–	–
PED residents (total)	–	63.96	2.3	63.88	62–66
EM resident (1)	38	–	–	–	–
EM resident (2)	40.5	–	–	–	–
EM resident (3)	40.5	–	–	–	–
EM resident (4)	43.75	–	–	–	–
EM resident (5)	46.25	–	–	–	–
EM resident (6)	49	–	–	–	–
EM resident (7)	51.75	–	–	–	–
EM residents (total)	–	44.25	4.64	43.75	41–49
ChatGPT-4o	72.5	–	–	–	–
Gemini 1.5 Pro	62.75	–	–	–	–
ChatGPT-4o mini	56.75	–	–	–	–
Gemini 1.5 Flash	56.5	–	–	–	–
Llama-3-8B	33.75	–	–	–	–
Chatbots (total)	–	56.45	12.76	56.75	57–63

As regards chatbots, Figure 5.4 allows to visually compare their scores through a bar plot. It also shows that significant difference is found between the total accuracy performance of ChatGPT-4o and ChatGPT-4o mini ($p<0.01$), and between ChatGPT-4o and Gemini 1.5 Flash ($p<0.01$). Llama-3-8B performes worse than all the other chatbots ($p<0.01$). No difference is observed between ChatGPT-4o and Gemini 1.5 Pro ($p=0.26$).

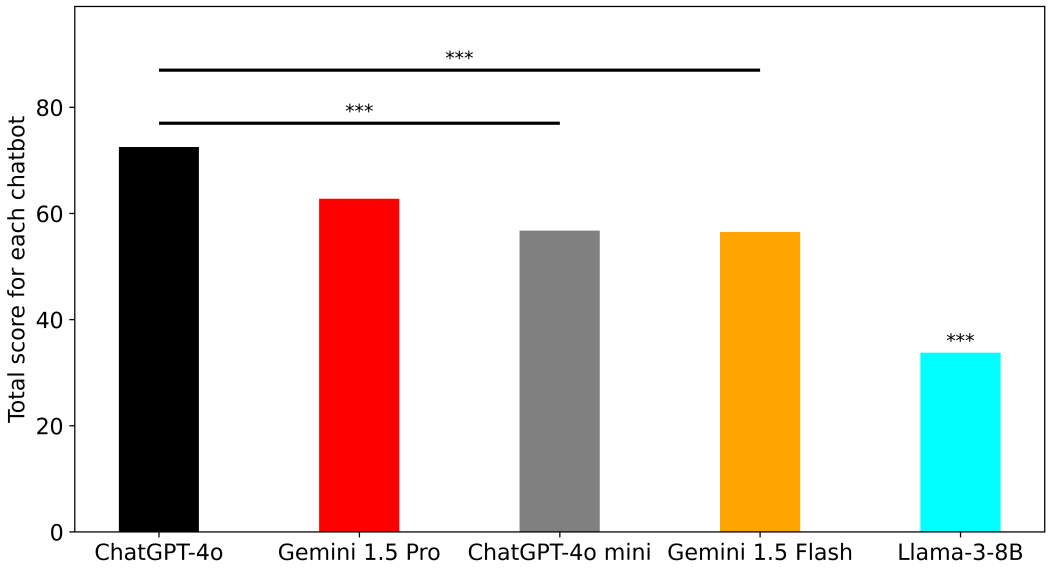


Fig. 5.4 Total scores of chatbots. The *** above the bar shows the p-values of the comparisons of that subject vs. all others. ***: $p<0.01$ (adapted from [154]).

Figure 5.5 compares the median total scores of physicians to the single performance of the best performing chatbots (ChatGPT-4o and Gemini 1.5 Pro). No significant difference is observed between PED residents and chatbots. However, ChatGPT-4o performs better than PED physicians ($p<0.05$), while EM residents perform worse than both the other physician groups and chatbots ($p<0.01$).

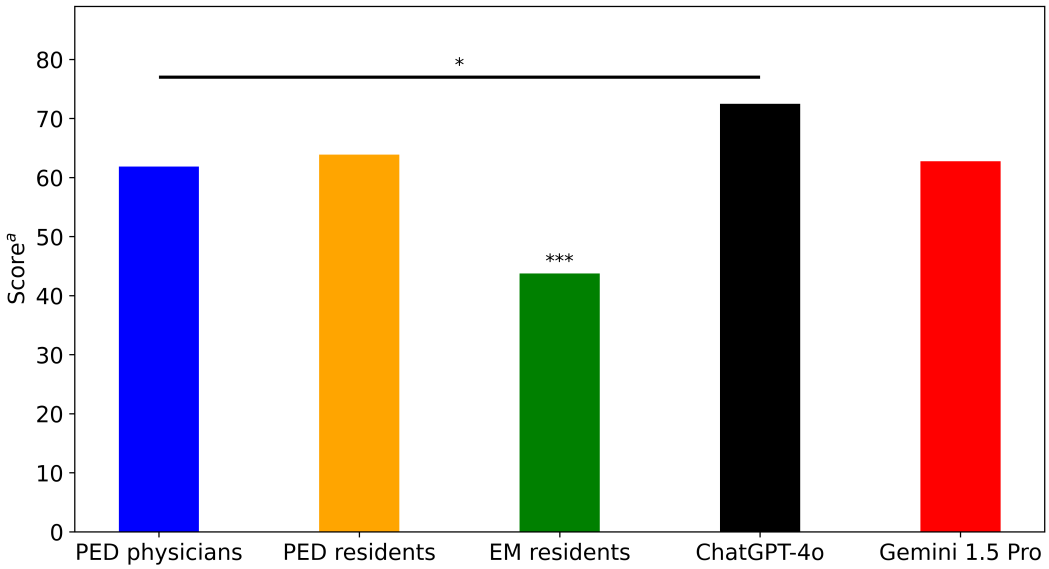
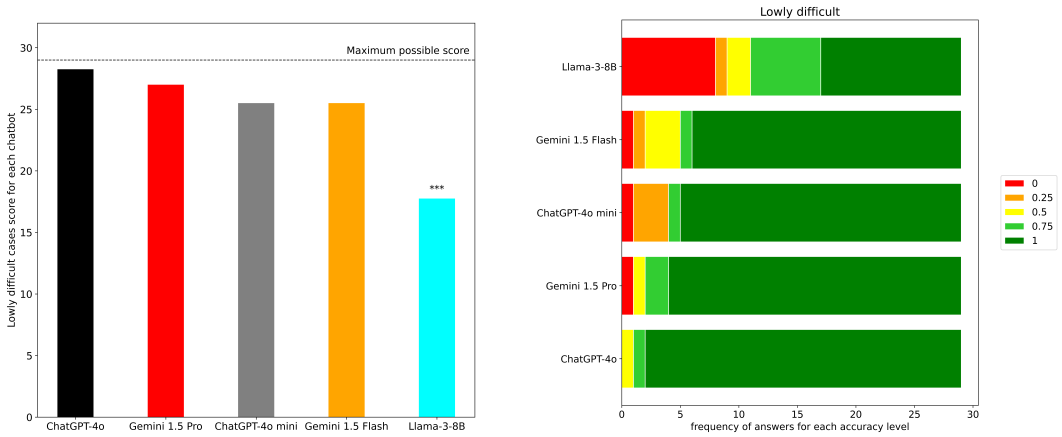


Fig. 5.5 Total scores of chatbots and physician subgroups. PED: pediatric emergency department; EM: emergency medicine. [a]: Median of total score for physicians. The *** above the bar shows the p-values of the comparisons of that subject vs. all others. *: p<0.05. ***: p<0.01 (adapted from [154]).

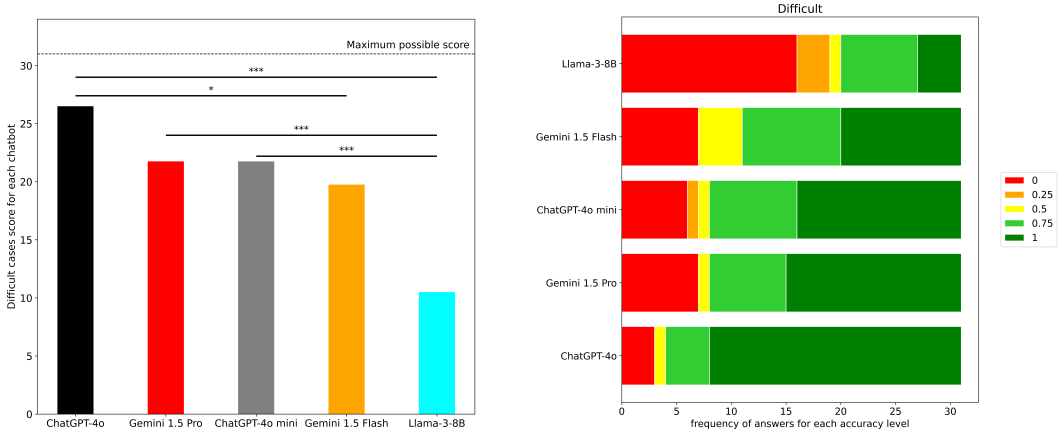
Afterwards, chatbots' diagnostic performance by case difficulty is analyzed, i.e., considering separately lowly difficult cases, difficult cases, and highly difficult cases. Table 5.5 reports the number of responses provided by each chatbot for each combination of accuracy level and difficulty; this information is then aggregated and illustrated in Figure 5.6. In particular, panels on the left show the total scores for each chatbot, while panels on the right show the frequency achieved for each of the 5 accuracy levels. In lowly difficult cases, all chatbots but Llama-3-8B perform well; Llama-3-8B shows a significant difference compared to other chatbots ($p<0.01$). In difficult cases, ChatGPT-4o performs better than Gemini 1.5 Flash ($p<0.05$) and Llama-3-8B ($p<0.01$). Gemini 1.5 Pro and ChatGPT-4o mini perform better than Llama-3-8B ($p<0.01$). No significant difference is found between ChatGPT-4o and Gemini 1.5 Pro, and between Gemini 1.5 Flash and Llama-3-8B. As regards highly difficult cases, ChatGPT-4o performs significantly better than ChatGPT-4o mini ($p<0.01$) and Llama-3-8B ($p<0.01$); also Gemini 1.5 Pro performs significantly better than Llama-3-8B ($p<0.01$). Importantly, ChatGPT-4o shows not only higher performance, but also better accuracy (Figure 5.6c), providing completely incorrect answers only in 4/80 cases (3 difficult, 1 highly difficult).

Table 5.5 Number of answers provided by each chatbot for each combination of accuracy level and difficulty. Lowly difficult cases: 29 (36.25%); difficult cases: 31 (38.75%); highly difficult cases: 20 (25.00%).

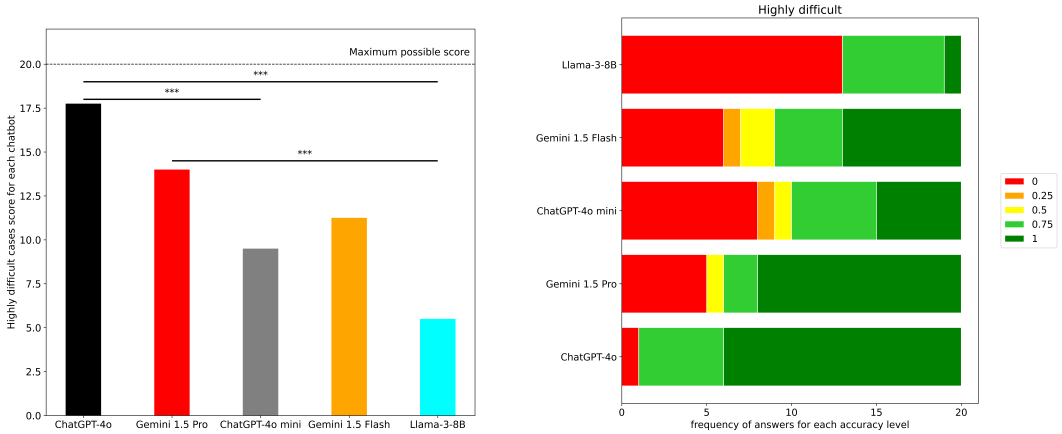
Accuracy level	Difficulty	ChatGPT-4o	Gemini 1.5 Pro	ChatGPT-4o mini	Gemini 1.5 Flash	Llama-3-8B
1	Lowly difficult	27	25	24	23	12
	Difficult	23	16	15	11	4
	Highly difficult	14	12	5	7	1
0.75	Lowly difficult	1	2	1	1	6
	Difficult	4	7	8	9	7
	Highly difficult	5	2	5	4	6
0.5	Lowly difficult	1	1	0	3	2
	Difficult	1	1	1	4	1
	Highly difficult	0	1	1	2	0
0.25	Lowly difficult	0	0	3	1	1
	Difficult	0	0	1	0	3
	Highly difficult	0	0	1	1	0
0	Lowly difficult	0	1	1	1	8
	Difficult	3	7	6	7	16
	Highly difficult	1	5	8	6	13



(a) Lowly difficult cases



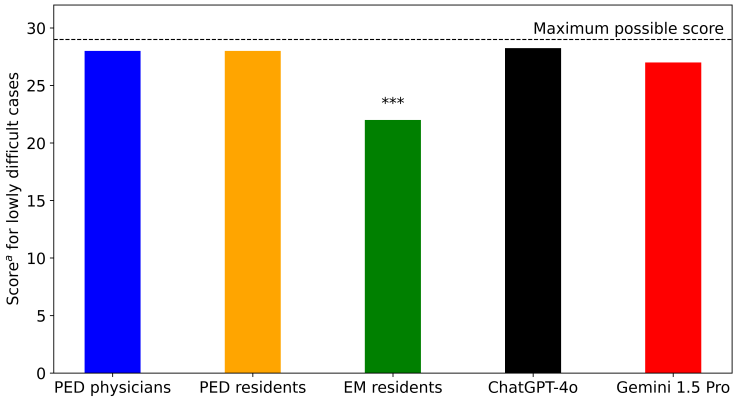
(b) Difficult cases



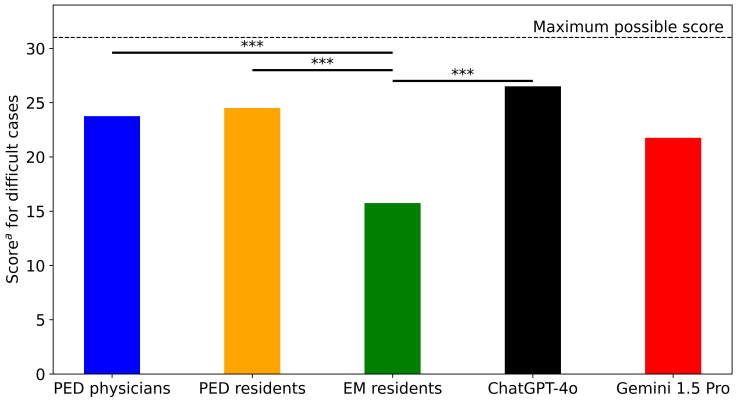
(c) Highly difficult cases

Fig. 5.6 Chatbots' diagnostic performance by case difficulty. Panels on the left show the total scores for each chatbot; panels on the right show the frequency of accuracy levels achieved. 5.6a Lowly difficult cases; 5.6b difficult cases; 5.6c highly difficult cases. The dashed line shows the maximum obtainable total score for the specific difficulty level. *: p-value<0.05. ***: p-value<0.01 (adapted from [154]).

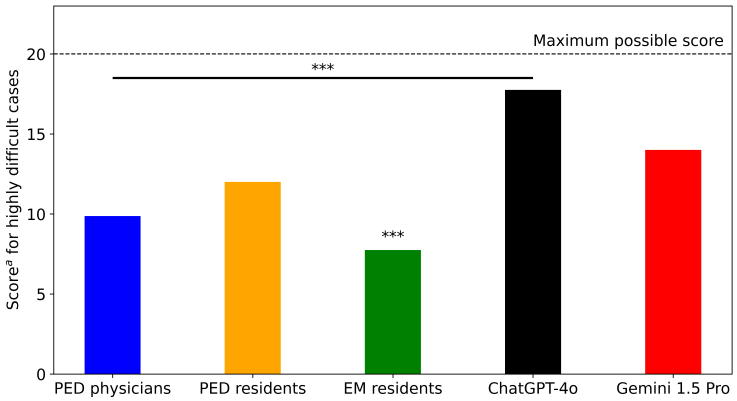
Last, the scores of the two best performing chatbots (ChatGPT-4o and Gemini 1.5 Pro) are compared to the median score obtained from the subgroups of physicians, stratified by difficulty (Figure 5.7). As regards the lowly and highly difficult cases, PED physicians, PED residents and both chatbots perform significantly better than EM residents ($p<0.01$). In difficult cases, PED physicians, PED residents and ChatGPT-4o perform significantly better than EM residents ($p<0.01$), but not Gemini 1.5 Pro ($p>0.05$). In highly difficult cases, both ChatGPT-4o and Gemini 1.5 Pro perform better than PED physicians and PED residents; however, statistical significance is reached only in the comparison between ChatGPT-4o and PED physicians ($p<0.01$).



(a) Lowly difficult cases



(b) Difficult cases



(c) Highly difficult cases

Fig. 5.7 Score of the best performing chatbots (ChatGPT-4o and Gemini 1.5 Pro) compared to the median score obtained from physician subgroups, stratified by case difficulty. 5.7a Lowly difficult cases; 5.7b difficult cases; 5.7c highly difficult cases. PED: pediatric emergency department; EM: emergency medicine. The dashed line shows the maximum obtainable score for the specific difficulty level. *: p-value<0.05. ***: p-value<0.01 (adapted from [154]).

5.5 Discussion

To our knowledge, this is the first study exploring the role of LLMs as diagnostic support tools in pediatric emergency cases. Among the tested chatbots, ChatGPT-4o achieves the highest accuracy, with most diagnoses aligning with correct answers for any level of complexity. In fact, ChatGPT provides a completely incorrect answer (i.e., scoring 0) in only 4 cases out of 80 (3 classified as difficult, and 1 as highly difficult). Gemini 1.5 Pro performs slightly below ChatGPT-4o, being more affected by case difficulty. Gemini 1.5 Flash and ChatGPT-4o mini achieve similar performance, but are inferior to ChatGPT-4o and Gemini 1.5 Pro: their performance is notably better in simpler cases, while it drops in difficult and highly difficult cases. In contrast, Llama-3-8B shows significantly lower performance than all the other LLMs considered in this research. This is aligned with expectations, as Llama-3-8B has only 8 billion parameters and the lowest scores on benchmarks [155, 189] and leaderboards [190]. However, during the study period, models like Gemini 1.5 Pro and ChatGPT-4o were paid services, with free questions available up to a daily limit; this may represent a limitation for some users. On the other hand, Llama-3-8B is open-source, free, and offers greater data privacy when used on-premises, though it requires a more complex setup and adequate computational resources compared to web-based chatbots. With more computational available resources, larger models such as Llama-3-70B [187] could be tested, offering significantly more parameters and potentially better performance.

Ultimately, this study underscores the importance of human oversight in the use of LLMs, as their success in healthcare stands on accurate data collection (e.g., medical history, physical examinations and vital signs) and interpretation, which only qualified practitioners can provide. LLMs are designed to complement physicians [194], whose role is not replaceable by AI since clinical data must be evaluated by a human and then be presented to AI in the correct way, such as in terms of language, in order to be analyzed effectively and usefully. Establishing specific clinical guidelines and protocols for the use of AI in healthcare is crucial to ensure in the future the safe integration of these tools into clinical practice. Looking ahead, the integration of LLMs into PED workflows such as electronic health records or diagnostic decision support systems is a desirable goal, but remains premature at this stage. Further research is needed to assess their reliability, clinical utility, and safe implementation in real-time diagnostic settings.

Regarding the physician groups, no significant difference in diagnostic accuracy is found in this study between PED physicians and PED residents, while a clear difference emerges between EM residents and the two pediatric physician groups. As expected, all the human subgroups show a decline in diagnostic accuracy as case complexity increases. ChatGPT-4o and Gemini 1.5 Pro perform like PED physicians and PED residents in lowly difficult and difficult cases, and prove to be effective aids in solving highly difficult cases (e.g., rare, complex diseases).

Interestingly, ChatGPT-4o performs better than both PED residents and PED physicians, but significance is reached only vs. the latter, particularly in highly difficult cases. This observation is difficult to interpret and could be due to different physician's subgroups sample size. It can be argued that PED residents perform better than PED physicians in those cases requiring knowledge of rare internal conditions, due to their more recent training. Anyway, the obtained results cannot support this hypothesis and further investigation on a larger sample should be carried out.

All LLMs outperform EM residents, likely due to their limited experience with pediatrics cases. In situations where a pediatrician is not immediately available, EM physicians could leverage the insights provided by LLMs alongside their own knowledge, allowing for initial diagnostic hypotheses. In the fast-paced ED environment, this could be a valuable advantage, speeding up the diagnostic process. On the other hand, this observation highlights the importance of implementing pediatric skills for EM residents, as in many cases children accessing the EDs are first evaluated by adult EM specialists, and not by specifically trained pediatricians. Pediatric skills should be not only acquired, but also maintained through longitudinal training programs during residency, as recently proposed [195].

While this study demonstrates the effectiveness of advanced LLMs in pediatric cases, a similar study by Barile et al. [196] showed significantly poorer outcomes using ChatGPT-3.5. Their investigation on 100 pediatric case challenges found a diagnostic error rate of 83%, highlighting limitations of older LLM versions. In contrast, our results indicate that state-of-art models (i.e., ChatGPT-4o and Gemini 1.5 Pro) achieve diagnostic accuracy comparable or even better than emergency pediatricians. This observation underscores the rapid advances in LLM technology and the importance of leveraging the most up-to-date tools to maximize clinical usefulness.

In fact, a general limitation when trying to evaluate LLMs performance in each context is the rapid advancement of these technologies, which can quickly make the results outdated. Moreover, LLMs are limited by the point in time when their training data are updated. If they are not fine-tuned or updated periodically, they may lack awareness of more recent data and information.

This study has some strengths. First, the effectiveness of the latest available versions of LLMs is evaluated, ranked among the top models on the Chatbot Arena leaderboard [190] and across various benchmarks [155, 189] at the time of the study. Such chatbots differ in model size, provider, user-interface, and availability. In contrast, many previous studies have focused on a single model, often an earlier version of ChatGPT [165–167, 196]. Second, three distinct groups of physicians are considered, allowing for diverse perspectives and detailed insights in addressing the assigned tasks. Last, a non-binary evaluation approach is introduced, using multiple accuracy categories to allow for more nuanced assessments.

The study also has some limits. First, as LLMs may show a lack of reproducibility, they could produce different responses when presented with the same case multiple times, sometimes reversing the order of diagnoses. This issue is not explored in this research.

Second, to avoid potential learning or contamination effects across prompt repetitions, each vignette was submitted only once per LLM. However, this approach prevents the assessment of intra-model variability. Future work should include repeated sampling to better quantify the consistency and stability of LLM-generated outputs. Sequential inputs or follow-up questions could also be explored, to simulate more closely real clinical conversations and evaluate their impact on diagnostic reasoning performance.

Moreover, when analyzing the physicians' responses, factors like a distracting environment, focus level, and stress or fatigue are not considered, which may increase inaccuracies, especially at the end of the forms. On the other hand, the process of reasoning on a clinical vignette is different from reasoning in front of a real patient: the clinical impression "at first sight" is crucial to reach the correct diagnosis and could be difficult to reproduce by written description [197]. Such limitations do not affect the responses provided by chatbots.

Furthermore, the varying number of cases across difficulty levels, with only 20 cases for the hardest ones, represents a limitation. Another limit is the non-

homogeneity of the number of physicians per group. This may affect the reliability of estimates for smaller groups, as they are more sensitive to outliers. In statistical analysis, a physician group's performance is summarized using the median score and compared with the absolute score of each chatbot. This difference in measurement may limit the accuracy of direct comparisons, affecting the generalizability of the results.

5.6 Conclusion

In conclusion, the results of this pilot study highlight the importance of understanding the diagnostic performance among different LLMs, especially in more complex PED clinical cases. The derived observations suggest that certain LLMs, especially ChatGPT-4o and Gemini 1.5 Pro, have diagnostic efficacy similar to or even better than those of pediatricians. Due to their high level of accuracy, LLMs could serve as a valuable tool to support PED physicians in solving the most difficult pediatric emergency cases, and they can be a very useful tool for EM physicians for all degrees of difficulty of pediatric cases. However, LLMs should never substitute human clinical judgement.

Chapter 6

Final Considerations and Future Directions

6.1 Main Contributions

This thesis investigates the potential of artificial intelligence, specifically computer vision and deep learning, to support diagnosis and clinical decision-making across different stages of human development. The target populations considered are heterogeneous, spanning older adults with cognitive disorders as well as newborns and young children who cannot verbally express their condition. This choice underscores the versatility and actual potential of the selected methodologies in addressing complex, heterogeneous challenges. This process is strengthened by the close multidisciplinary collaboration with healthcare professionals.

Overall, the work presented in this thesis introduces a set of original contributions that advance both the technical and clinical understanding of AI-assisted healthcare. These contributions can be summarized as follows.

1. *Cognitive impairment detection.* An AI-based framework is developed to identify cognitive impairment through the automatic analysis of facial emotional responses. The system achieves solid performance in distinguishing cognitively impaired individuals from healthy subjects, even at early stages, and in differentiating Alzheimer's disease from other forms of cognitive decline.

These results demonstrate the potential of this approach as a non-invasive tool to support early and differential dementia diagnosis.

2. *Automatic pain assessment.* An automated system for pain assessment in children under three years of age is developed and tested in a real Pediatric Emergency Department setting, using behavioral features extracted from video recordings. Moreover, a deep learning approach relying solely on facial expressions is designed for newborn pain detection and validated on a newly collected dataset from the Neonatology Department, demonstrating its feasibility in real clinical conditions. Explainable AI techniques are also integrated to highlight the facial regions most relevant for pain detection, increasing model transparency and supporting clinical trust.

3. *Large Language Model diagnostic performance.* A framework is developed to assess the diagnostic capabilities of Large Language Models in the Pediatric Emergency Department. The performance of five widely used models is evaluated on real pediatric emergency cases and compared with that of a group of physicians. Since some models (ChatGPT-4o and Gemini 1.5 Pro) achieve diagnostic accuracy comparable to, or in some cases exceeding, that of clinicians, their potential as supportive tools in emergency pediatric care is highlighted.

6.2 Overall Challenges and Limitations

Across all studies included in this thesis, several challenges and limitations have emerged during the development, implementation, and evaluation of AI methodologies in real clinical contexts. These issues arise from both the technical nature of AI systems and the intrinsic complexity of clinical environments, and they significantly influence the interpretation and generalizability of the obtained results.

From a technical perspective, one of the strongest constraints concerns the size, quality, and diversity of available data. First, video acquisition in hospital settings is inevitably affected by variations in patient positioning, movement, lighting, and camera operation. These factors may compromise the performance of face detection and pose estimation algorithms, which are highly sensitive to such variations. Second, recruiting vulnerable populations such as newborns, young children, or patients with

cognitive impairment is particularly challenging. Ethical, logistical, and practical considerations often limit the number of participants, and therefore the size and diversity of the collected datasets. Limited sample sizes and class imbalances, as in the dementia study or in newborn pain detection, may hinder the development of models capable of capturing the full variability of facial expressions and clinical presentations, and inevitably impact model robustness. As a consequence, generalization remains a major limitation: systems performing well on a specific dataset or in a controlled setting may not transfer reliably to different hospitals, imaging devices, or demographic groups.

Methodologically, these constraints require the adoption of strategies specifically designed for small datasets, including careful model design, appropriate cross-validation schemes, and domain-specific preprocessing. However, such strategies cannot fully compensate for the absence of large, diverse datasets, and their effectiveness diminishes in the presence of highly heterogeneous or movement-intensive data. For these reasons, substantial efforts toward collecting larger and more diverse datasets are essential, particularly if the ultimate goal is to develop AI systems that can be reliably deployed in real clinical environments.

A further important challenge concerns model interpretability. For AI systems to be meaningfully adopted in clinical practice, healthcare professionals must be able to understand the factors driving model decisions and correctly interpret their outputs. Black-box models, which offer limited insight into the underlying decision process, hinder clinical trust and make it difficult to integrate automated assessments into existing diagnostic workflows.

Across the different studies included in this thesis, several steps are taken to address this issue. In cognitive impairment detection, interpretability is supported by extracting emotional features from facial expressions, which clinicians have found more intuitive and clinically meaningful than using raw images as model inputs. For young children, the automatic pain-detection system directly maps its outputs to the traditional pain-scale scores used in clinical practice, making the results easier to understand and compare with human evaluations. In newborn pain detection, explainable AI techniques are applied more explicitly: Grad-CAM visualizations highlight the facial regions most influential for the model's decisions, thereby improving transparency and helping clinicians assess whether the automated reasoning aligns with their own observations.

Despite these efforts, much remains to be explored. Recent advances in explainable AI offer more sophisticated strategies capable of revealing expressive, behavioral, or temporal features that drive model predictions, even in end-to-end deep learning architectures. Incorporating these approaches could significantly enhance clinicians' ability to interpret automated assessments and increase confidence in their use in real clinical environments.

Finally, this work is characterized also by clinical and operational challenges. Data collection and processing must comply with strict ethical, privacy, and regulatory requirements. Moreover, the cross-disciplinary nature of this research requires constant dialogue between technical and clinical teams. Defining meaningful clinical objectives, interpreting model outputs, and validating results all depend on effective collaboration between engineers and healthcare professionals. Differences in language, expectations, and methodological frameworks can slow this process and often require iterative adjustments to both computational models and clinical protocols. However, the collaborative and interdisciplinary approach adopted throughout this thesis has proven essential to ensuring that the proposed systems remain clinically relevant, and aligned with real healthcare needs. At the same time, the challenges highlighted here point to clear opportunities for future methodological refinements and practical improvements, which are discussed in the following section.

6.3 Future Perspectives

Looking ahead, several directions emerge as promising avenues to further strengthen the methodological robustness of the proposed systems and support their eventual integration into clinical practice. A priority for future research is the substantial expansion of datasets across all clinical domains examined in this thesis. Multicenter data collection encompassing diverse ethnicities, hospitals, and clinical settings will be essential to assess and improve the generalizability of AI models. In the dementia study, larger and more heterogeneous cohorts would enable the exploration of subtler disease subtypes. Similarly, in pediatric and neonatal contexts, increasing the sample size would support finer-grained analyses, such as multi-level pain intensity classification, and allow a more comprehensive characterization of behavioral variability.

Beyond data collection, several methodological enhancements could further improve technical robustness and support future deployment. Multimodal approaches integrating facial expressions with physiological signals (e.g., heart rate, skin conductance, skin temperature) hold particular promise for more reliable and clinically comprehensive assessments. Incorporating temporal information from video sequences also represents a crucial step toward real-time pain monitoring and richer modeling of dynamic behavioral patterns. Furthermore, improving model interpretability remains a central objective. Enhanced transparency not only increases clinician trust but also facilitates meaningful clinical validation.

Finally, moving towards real-world deployment requires careful attention to operational and clinical integration. Future work should consider the development of user-friendly interfaces, evaluation of real-time feasibility, and alignment with regulatory frameworks. Establishing clinical protocols in which automated assessments complement, rather than replace, human expertise will be essential to ensure safe and effective adoption.

References

[1] Letizia Bergamasco, Federica Lorenzo, Anita Coletta, Gabriella Olmo, Aurora Cermelli, Elisa Rubino, and Innocenzo Rainero. Automatic detection of cognitive impairment through facial emotion analysis. *Applied Sciences*, 15(16):9103, 2025.

[2] Letizia Bergamasco, Anita Coletta, Gabriella Olmo, Aurora Cermelli, Elisa Rubino, and Innocenzo Rainero. AI-based facial emotion analysis for early and differential diagnosis of dementia. *Bioengineering*, 12(10):1082, 2025.

[3] World Health Organization. *Global status report on the public health response to dementia*. 2021.

[4] Philip Scheltens, Kaj Blennow, Monique MB Breteler, Bart De Strooper, Giovanni B Frisoni, Stephen Salloway, and Wiesje Maria Van der Flier. Alzheimer's disease. *The Lancet*, 388(10043):505–517, 2016.

[5] John T O'Brien and Alan Thomas. Vascular dementia. *The Lancet*, 386(10004):1698–1706, 2015.

[6] Jee Bang, Salvatore Spina, and Bruce L Miller. Frontotemporal dementia. *The Lancet*, 386(10004):1672–1682, 2015.

[7] Zuzana Walker, Katherine L Possin, Bradley F Boeve, and Dag Aarsland. Lewy body dementias. *The Lancet*, 386(10004):1683–1697, 2015.

[8] Alain Koyama, Olivia I Okereke, Ting Yang, Deborah Blacker, Dennis J Selkoe, and Francine Grodstein. Plasma amyloid-β as a predictor of dementia and cognitive decline: a systematic review and meta-analysis. *Archives of neurology*, 69(7):824–831, 2012.

[9] G Caleb Alexander, Scott Emerson, and Aaron S Kesselheim. Evaluation of aducanumab for alzheimer disease: scientific evidence and regulatory review involving efficacy, safety, and futility. *Jama*, 325(17):1717–1718, 2021.

[10] Christopher H Van Dyck, Chad J Swanson, Paul Aisen, Randall J Bateman, Christopher Chen, Michelle Gee, Michio Kanekiyo, David Li, Larisa Reyderman, Sharon Cohen, et al. Lecanemab in early alzheimer's disease. *New England Journal of Medicine*, 388(1):9–21, 2023.

[11] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.

[12] Kuan-Hua Chen, Sandy J Lwi, Alice Y Hua, Claudia M Haase, Bruce L Miller, and Robert W Levenson. Increased subjective experience of non-target emotions in patients with frontotemporal dementia and alzheimer's disease. *Current Opinion in Behavioral Sciences*, 15:77–84, 2017.

[13] Peter S Pressman, Kuan Hua Chen, James Casey, Stefan Sillau, Heidi J Chial, Christopher M Filley, Bruce L Miller, and Robert W Levenson. Incongruences between facial expression and self-reported emotional reactivity in frontotemporal dementia and related disorders. *The Journal of neuropsychiatry and clinical neurosciences*, 35(2):192–201, 2023.

[14] Jian Sun, Hiroko H Dodge, and Mohammad H Mahoor. MC-ViViT: Multi-branch classifier-ViViT to detect mild cognitive impairment in older adults using facial videos. *Expert Systems with Applications*, 238:121929, 2024.

[15] 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):1765–1772, 2021.

[16] Zixiang Fei, Erfu Yang, Leijian Yu, Xia Li, Huiyu Zhou, and Wenju Zhou. A novel deep neural network-based emotion analysis system for automatic detection of mild cognitive impairment in the elderly. *Neurocomputing*, 468:306–316, 2022.

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

[18] Hiroko H Dodge, Kexin Yu, Chao-Yi Wu, Patrick J Pruitt, Meysam Asgari, Jeffrey A Kaye, Benjamin M Hampstead, Laura Struble, Kathleen Potempa, Peter Lichtenberg, Raina Croff, Roger L Albin, Lisa C Silbert, and the I-CONECT Team. 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. *The Gerontologist*, 64(4):gnad147, 11 2023.

[19] 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), 2023.

[20] Taishiro Kishimoto, Akihiro Takamiya, Kuo-ching Liang, Kei Funaki, Takanori Fujita, Momoko Kitazawa, Michitaka Yoshimura, Yuki Tazawa, Toshiro Horigome, Yoko Eguchi, Toshiaki Kikuchi, Masayuki Tomita, Shogyoku Bun, Junichi Murakami, Brian Sumali, Tifani Warnita, Aiko Kishi, Mizuki Yotsui, Hiroyoshi Toyoshiba, Yasue Mitsukura, Koichi Shinoda, Yasubumi Sakakibara, and Masaru Mimura. The project for objective measures using computational psychiatry technology (PROMPT): Rationale, design, and methodology. *Contemporary Clinical Trials Communications*, 19:100649, 2020.

[21] Zifan Jiang, Salman Seyedi, Rafi U Haque, Alvince L Pongos, Kayci L Vickers, Cecelia M Manzanares, James J Lah, Allan I Levey, and Gari D Clifford. Automated analysis of facial emotions in subjects with cognitive impairment. *Plos one*, 17(1):e0262527, 2022.

[22] Taichi Okunishi, Chuheng Zheng, Mondher Bouazizi, Tomoaki Ohtsuki, Momoko Kitazawa, Toshiro Horigome, and Taishiro Kishimoto. Dementia and mci detection based on comprehensive facial expression analysis from videos during conversation. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3537–3548, 2025.

[23] Che-Sheng Chu, Di-Yuan Wang, Chih-Kuang Liang, Ming-Yueh Chou, Ying-Hsin Hsu, Yu-Chun Wang, Mei-Chen Liao, Wei-Ta Chu, and Yu-Te Lin. Automated video analysis of audio-visual approaches to predict and detect mild cognitive impairment and dementia in older adults. *Journal of Alzheimer's Disease*, 92(3):875–886, 2023.

[24] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124–129, 1971.

[25] Robert Plutchik. A psycho evolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553, 1982.

[26] Machine Elf 1735. Plutchik's wheel of emotions. https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg, 2006. Public domain image, via Wikimedia Commons. Accessed October 27, 2025.

[27] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[28] Bryn Farnsworth. Facial action coding system (facs) – a visual guidebook. https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/, 2022. Accessed September 1, 2025.

[29] Vedant Chauhan, Yash Agrawal, and Vinay Bhutada. Emotion detection system using facial action coding system. *International Journal of Engineering and Technical Research*, pages 2321–0869, 2016.

[30] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[31] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[32] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101. IEEE, 2010.

[33] Lutfiah Zahara, Purnawarman Musa, Eri Prasetyo Wibowo, Irwan Karim, and Saiful Bahri Musa. The Facial Emotion Recognition (FER-2013) dataset for prediction system of micro-expressions face using the Convolutional Neural Network (CNN) algorithm based Raspberry Pi. In *2020 Fifth international conference on informatics and computing (ICIC)*, pages 1–9. IEEE, 2020.

[34] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[35] Michael J Lyons. "Excavating AI" re-excavated: debunking a fallacious account of the JAFFE dataset. *arXiv preprint arXiv:2107.13998v1*, 2021. Accessed October 27, 2025.

[36] Dimitrios Kollias and Stefanos Zafeiriou. Aff-Wild2: extending the Aff-Wild database for affect recognition. *arXiv preprint arXiv:1811.07770v2*, 2019. Accessed October 27, 2025.

[37] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[38] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The OMG-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[39] Liam Schoneveld and Alice Othmani. Towards a general deep feature extractor for facial expression recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2339–2342. IEEE, 2021.

[40] Shekhar Singh and Fatma Nasoz. Facial expression recognition with convolutional neural networks. In *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0324–0328. IEEE, 2020.

[41] Sebastian Handrich, Laslo Dinges, Ayoub Al-Hamadi, Philipp Werner, and Zaher Al Aghbari. Simultaneous prediction of valence/arousal and emotions

on AffectNet, Aff-Wild and AFEW-VA. *Procedia Computer Science*, 170:634–641, 2020.

[42] Thomas Teixeira, Eric Granger, and Alessandro Lameiras Koerich. Continuous emotion recognition with spatiotemporal convolutional neural networks. *Applied Sciences*, 11(24), 2021.

[43] Jing Li, Kan Jin, Dalin Zhou, Naoyuki Kubota, and Zhaojie Ju. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing*, 411:340–350, 2020.

[44] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 580:35–54, 2021.

[45] Wang Xiaohua, Peng Muzi, Pan Lijuan, Hu Min, Jin Chunhua, and Ren Fuji. Two-level attention with two-stage multi-task learning for facial emotion recognition. *Journal of Visual Communication and Image Representation*, 62:217–225, 2019.

[46] Mohan Karnati, Ayan Seal, Debotosh Bhattacharjee, Anis Yazidi, and Ondrej Krejcar. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Transactions on Instrumentation and Measurement*, 72:1–31, 2023.

[47] François Chollet. Keras. https://keras.io, 2015. Accessed January 29, 2025.

[48] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51:195–203, 2019.

[49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[50] Quan Ngo and Seokhoon Yoon. Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset. *Sensors*, 20:2639, 2020.

[51] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[52] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[53] keras vggface. VGGFace implementation with Keras framework. https://github.com/rcmalli/keras-vggface. Accessed January 29, 2025.

[54] Michelle Yik, Chiel Mues, Irene NL Sze, Peter Kuppens, Francis Tuerlinckx, Kim De Roover, Felity HC Kwok, Shalom H Schwartz, Maher Abu-Hilal, Damilola Fisayo Adebayo, et al. On the relationship between valence and arousal in samples across the globe. *Emotion*, 23(2):332–344, 2023.

[55] Rajesh Nandy, Karabi Nandy, and Scott T Walters. 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 Formative Research*, 7:e34989, 2023.

[56] Srinivas Parthasarathy and Carlos Busso. Jointly predicting arousal, valence and dominance with multi-task learning. In *Proceedings of the Interspeech 2017*, pages 1103–1107, 2017.

[57] Marko Horvat, Davor Kukolja, and Dragutin Ivanec. 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)*, pages 1394–1398. IEEE, 2015.

[58] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 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, 2008.

[59] Margaret M Bradley and Peter J Lang. 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, 2007.

[60] Kenneth Yuen, Stephen Johnston, Federico Martino, Bettina Sorger, Elia Formisano, David Linden, and Rainer Goebel. Pattern classification predicts individuals' responses to affective stimuli. *Translational Neuroscience*, 3(3):278–287, 2012.

[61] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis. Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Transactions on affective computing*, 1(2):81–97, 2010.

[62] Vrinda Prajapati, Rajlakshmi Guha, and Aurobinda Routray. Multimodal prediction of trait emotional intelligence-Through affective changes measured using non-contact based physiological measures. *Plos one*, 16(7):e0254335, 2021.

[63] Walied Merghani, Adrian K Davison, and Moi Hoon Yap. A review on facial micro-expressions analysis: Datasets, features and metrics. *arXiv preprint arXiv:1805.02397v1*, 2018. Accessed October 27, 2025.

[64] Clifford R Jack Jr, J Scott Andrews, Thomas G Beach, Teresa Buracchio, Billy Dunn, Ana Graf, Oskar Hansson, Carole Ho, William Jagust, Eric McDade, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's & Dementia*, 20(8):5143–5169, 2024.

[65] Giovanni B Frisoni, Cristina Festari, Federico Massa, Matteo Cotta Ramusino, Stefania Orini, Dag Aarsland, Federica Agosta, Claudio Babiloni, Barbara Borroni, Stefano F Cappa, et al. European intersocietal recommendations for the biomarker-based diagnosis of neurocognitive disorders. *The Lancet Neurology*, 23(3):302–312, 2024.

[66] Katya Rascovsky, John R Hodges, David Knopman, Mario F Mendez, Joel H Kramer, John Neuhaus, John C Van Swieten, Harro Seelaar, Elise GP Dopper, Chiadi U Onyike, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, 134(9):2456–2477, 2011.

[67] Ian G McKeith, Bradley F Boeve, Dennis W Dickson, Glenda Halliday, John-Paul Taylor, Daniel Weintraub, Dag Aarsland, James Galvin, Johannes Attems, Clive G Ballard, et al. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. *Neurology*, 89(1):88–100, 2017.

[68] Perminder Sachdev, Raj Kalaria, John O'Brien, Ingmar Skoog, Suvarna Alladi, Sandra E Black, Deborah Blacker, Dan G Blazer, Christopher Chen, Helena Chui, et al. Diagnostic criteria for vascular cognitive disorders: a VASCOG statement. *Alzheimer Disease & Associated Disorders*, 28(3):206–218, 2014.

[69] Stephen M Wilson, Sebastiano Galantucci, Maria Carmela Tartaglia, and Maria Luisa Gorno-Tempini. The neural basis of syntactic deficits in primary progressive aphasia. *Brain and Language*, 122(3):190–198, 2012.

[70] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. Accessed July 28, 2025.

[71] Ivan Grishchenko and Valentin Bazarevsky. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/. Accessed July 28, 2025.

[72] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2), 2023.

[73] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2):1236–1248, 2023.

[74] Enrico Pellegrini, Lucia Ballerini, Maria del C. Valdes Hernandez, Francesca M Chappell, Victor González-Castro, Devasuda Anblagan, Samuel Danso, Susana Muñoz-Maniega, Dominic Job, Cyril Pernet, Grant Mair, Tom J MacGillivray, Emanuele Trucco, and Joanna M Wardlaw. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:519–535, 2018.

[75] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[76] Empatica. Embraceplus. the world's most advanced smartwatch for continuous health monitoring. https://www.empatica.com/embraceplus/. Accessed October 20, 2025.

[77] Boxuan Zhong, Zikun Qin, Shuo Yang, Junyu Chen, Nicholas Mudrick, Michelle Taub, Roger Azevedo, and Edgar Lobaton. Emotion recognition with facial expressions and physiological signals. In *2017 IEEE symposium series on computational intelligence (SSCI)*, pages 1–8. IEEE, 2017.

[78] Chuan-Yu Chang, Jeng-Shiun Tsai, Chi-Jane Wang, and Pau-Choo Chung. Emotion recognition with consideration of facial expression and physiological signals. In *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 278–283. IEEE, 2009.

[79] Letizia Bergamasco, Marco Gavelli, Carla Fadda, Emilia Parodi, Claudia Bondone, and Emanuele Castagno. Measurement of acute pain in the pediatric emergency department through automatic detection of behavioral parameters: A pilot study. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 469–481. Springer Nature Switzerland, 2023.

[80] Paul S Grant. Analgesia delivery in the ED. *The American journal of emergency medicine*, 24(7):806–809, 2006.

[81] Srinivasa N Raja, Daniel B Carr, Milton Cohen, Nanna B Finnerup, Herta Flor, Stephen Gibson, Francis J Keefe, Jeffrey S Mogil, Matthias Ringkamp, Kathleen A Sluka, et al. The revised international association for the study of pain definition of pain: concepts, challenges, and compromises. *Pain*, 161(9):1976–1982, 2020.

[82] Suellen M Walker. Neonatal pain. *Pediatric Anesthesia*, 24(1):39–48, 2014.

[83] Kathryn A Birnie, Christine T Chambers, Conrad V Fernandez, Paula A Forgeron, Margot A Latimer, Patrick J McGrath, Elizabeth A Cummings, and G Allen Finley. Hospitalized children continue to report undertreated and preventable pain. *Pain Research and Management*, 19(4):198–204, 2014.

[84] Gunilla Brattberg. Do pain problems in young school children persist into early adulthood? a 13-year follow-up. *European Journal of Pain*, 8(3):187–199, 2004.

[85] Amy L Drendel, David C Brousseau, and Marc H Gorelick. Pain assessment for pediatric patients in the emergency department. *Pediatrics*, 117(5):1511–1518, 2006.

[86] Baruch S Krauss, Lorenzo Calligaris, Steven M Green, and Egidio Barbi. Current concepts in management of pain in children in the emergency department. *The Lancet*, 387(10013):83–92, 2016.

[87] Federico Marzona, Sara Pedicini, Eva Passone, Anna Pusiol, and Paola Cogo. Mandatory pain assessment in a pediatric emergency department: failure or success?: a retrospective study. *The Clinical Journal of Pain*, 35(10):826–830, 2019.

[88] Franca Benini, Ilaria Corsini, Emanuele Castagno, Davide Silvagni, Annunziata Lucarelli, Luca Giacomelli, Angela Amigoni, Gina Ancora, Marinella Astuto, Fabio Borrometi, et al. Consensus on pediatric pain in the emergency room: the COPPER project, issued by 17 italian scientific societies. *Italian journal of pediatrics*, 46(1):101, 2020.

[89] Sandra I Merkel, Terri Voepel-Lewis, Jay R Shayevitz, and Shobha Malviya. The FLACC: a behavioral scale for scoring postoperative pain in young children. *Pediatric Nursing*, 23(3):293–297, 1997.

[90] Ruth VE Grunau and Kenneth D Craig. Pain expression in neonates: facial action and cry. *Pain*, 28(3):395–410, 1987.

[91] Sheryl Brahnam, Chao-Fa Chuang, Frank Y. Shih, and Melinda R. Slack. Machine recognition and representation of neonatal facial displays of acute pain. *Artificial Intelligence in Medicine*, 36(3):211–222, 2006.

[92] Sheryl Brahnam, Loris Nanni, and Randall Sexton. Introduction to neonatal facial pain detection using common and advanced face classification techniques. In *Advanced Computational Intelligence Paradigms in Healthcare – 1*, pages 225–253. Springer Berlin Heidelberg, 2007.

[93] Ghada Zamzami, Gabriel Ruiz, Dmitry Goldgof, Rangachar Kasturi, Yu Sun, and Terri Ashmeade. Pain assessment in infants: Towards spotting pain expression based on infants' facial strain. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 5, pages 1–5. IEEE, 2015.

[94] Sun K Yoo, Chung K Lee, Youn J Park, Nam H Kim, Byung C Lee, and Kee S Jeong. Neural network based emotion estimation using heart rate variability and skin resistance. In *International conference on natural computation*, pages 818–824. Springer Berlin Heidelberg, 2005.

[95] Manon Ranger and Céline Gélinas. Innovating in pain assessment of the critically ill: exploring cerebral near-infrared spectroscopy as a bedside approach. *Pain Management Nursing*, 15(2):519–529, 2014.

[96] Hodjat Rahmati, Ole Morten Aamo, Øyvind Stavdahl, Ralf Dragon, and Lars Adde. Video-based early cerebral palsy prediction using motion segmentation. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3779–3783. IEEE, 2014.

[97] Annette Stahl, Christian Schellewald, Øyvind Stavdahl, Ole Morten Aamo, Lars Adde, and Harald Kirkerod. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(4):605–614, 2012.

[98] Ghada Zamzmi, Dmitry Goldgof, Rangachar Kasturi, Yu Sun, and Terri Ashmeade. Machine-based multimodal pain assessment tool for infants: A review. *arXiv preprint arXiv:1607.00331*, 2019. Accessed July 28, 2025.

[99] Emilia Parodi, Daniela Melis, Luca Boulard, Marco Gavelli, and Enrico Baccaglini. Automated newborn pain assessment framework using computer vision techniques. In *Proceedings of the 4th International Conference on Bioinformatics Research and Applications*, page 31–36. Association for Computing Machinery, 2017.

[100] Alessandra Di Bari, Anne Destrebecq, Federica Osnaghi, and Stefano Terzoni. Traduzione e validazione in italiano della scala Revised FLACC per la valutazione del dolore nel bambino con grave ritardo mentale. *Pain Nursing Magazine - Italian Online Journal*, 4-2013, 2013.

[101] Abirami Vina. How to use ultralytics yolo11 for pose estimation. https://www.ultralytics.com/blog/how-to-use-ultralytics-yolo11-for-pose-estimation. Accessed November 26, 2025.

[102] Sebastian Janampa and Marios Pattichis. Detrpose: Real-time end-to-end transformer model for multi-person pose estimation. *arXiv preprint arXiv:2506.13027*, 2025. Accessed November 26, 2025.

[103] Jure Kovac, Peter Peer, and Franc Solina. Human skin color clustering for face detection. In *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, volume 2, pages 144–148. IEEE, 2003.

[104] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1859–1866, 2014.

[105] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[106] Letizia Bergamasco, Marta Lattanzi, Marco Gavelli, Claudio Pastrone, Gabriella Olmo, Lucia Borsotti, and Emilia Parodi. Pain assessment in neonatal clinical practice via facial expression analysis and deep learning. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 249–263. Springer Nature Switzerland, 2024.

[107] Carlo Valerio Bellieni. Pain assessment in human fetus and infants. *The AAPS journal*, 14:456–461, 2012.

[108] Ruth E Grunau, Liisa Holsti, David W Haley, Tim Oberlander, Joanne Weinberg, Alfonso Solimano, Michael F Whitfield, Colleen Fitzgerald, and Wayne Yu. Neonatal procedural pain exposure predicts lower cortisol and behavioral reactivity in preterm infants in the nicu. *Pain*, 113(3):293–300, 2005.

[109] Anna Taddio and Joel Katz. The effects of early pain experience in neonates on pain responses in infancy and childhood. *Pediatric Drugs*, 7:245–257, 2005.

[110] Suellen M Walker. Biological and neurodevelopmental implications of neonatal pain. *Clinics in Perinatology*, 40(3):471–491, 2013.

[111] Emma Olsson, Hanna Ahl, Kevin Bengtsson, Dhashini N Vejayaram, Elisabeth Norman, Matteo Bruschettini, and Mats Eriksson. The use and reporting of neonatal pain scales: a systematic review of randomized trials. *Pain*, 162(2):353–360, 2021.

[112] Bonnie Stevens, Celeste Johnston, Patricia Petryshen, and Anna Taddio. Premature infant pain profile: development and initial validation. *The Clinical journal of pain*, 12(1):13–22, 1996.

[113] Ricardo Carbajal, Alain Paupe, E. Hoenn, Richard Lenclen, and Marie Olivier-Martin. DAN : une échelle comportementale d'évaluation de la douleur aiguë du nouveau-né. *Archives de Pédiatrie*, 4(7):623–628, 1997.

[114] Md Sirajus Salekin, Ghada Zamzmi, Rahul Paul, Dmitry Goldgof, Rangachar Kasturi, Thao Ho, and Yu Sun. Harnessing the power of deep learning methods in healthcare: Neonatal pain assessment from crying sound. In *2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT)*, pages 127–130. IEEE, 2019.

[115] Vito Giordano, Alexandra Luister, Christoph Reuter, Isabella Czedik-Eysenberg, Dominique Singer, David Steyrl, Eik Vettorazzi, and Philipp Deindl. Audio feature analysis for acoustic pain detection in term newborns. *Neonatology*, 119(6):760–768, 2022.

[116] Tatiany M Heiderich, Lucas P Carlini, Lucas F Buzuti, Rita de CX Balda, Marina CM Barros, Ruth Guinsburg, and Carlos E Thomaz. Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units. *Jornal de Pediatria*, 99(6):546–560, 2023.

[117] Huaiyu Zhu, Yisheng Zhao, Xiaofei Chen, Feixiang Luo, Lingli Mei, Shuohui Chen, and Yun Pan. Video-based neonatal pain assessment in uncontrolled conditions. *IEEE Journal of Biomedical and Health Informatics*, 28(1):239–250, 2024.

[118] Md Sirajus Salekin, Ghada Zamzmi, Dmitry Goldgof, Peter R Mouton, Kanwaljeet JS Anand, Terri Ashmeade, Stephanie Prescott, Yangxin Huang, and Yu Sun. Attentional generative multimodal network for neonatal postoperative pain estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 749–759. Springer Nature Switzerland, 2022.

[119] Sheryl Brahnam, Chao-Fa Chuang, Frank Y Shih, and Melinda R Slack. Svm classification of neonatal facial images of pain. In *International workshop on fuzzy logic and applications*, pages 121–128. Springer Berlin Heidelberg, 2005.

[120] Sheryl Brahnam, Chao-Fa Chuang, Randall S Sexton, and Frank Y Shih. Machine assessment of neonatal facial expressions of acute pain. *Decision Support Systems*, 43(4):1242–1254, 2007.

[121] Sheryl Brahnam, Loris Nanni, Shannon McMurtrey, Alessandra Lumini, Rick Brattin, Melinda Slack, and Tonya Barrier. Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from gaussian of local descriptors. *Applied Computing and Informatics*, 19(1/2):122–143, 2023.

[122] Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof, Rangachar Kasturi, Terri Ashmeade, and Yu Sun. A comprehensive and context-sensitive neonatal pain assessment using computer vision. *IEEE Transactions on Affective Computing*, 13(1):28–45, 2022.

[123] Md Sirajus Salekin, Ghada Zamzmi, Jacqueline Hausmann, Dmitry Goldgof, Rangachar Kasturi, Marcia Kneusel, Terri Ashmeade, Thao Ho, and Yu Sun. Multimodal neonatal procedural and postoperative pain assessment dataset. *Data in Brief*, 35:106796, 2021.

[124] Joy Egede, Michel Valstar, Mercedes Torres Torres, and Don Sharkey. Automatic neonatal pain estimation: An acute pain in neonates database. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[125] Jingjie Yan, Guanming Lu, Xiaonan Li, Wenming Zheng, Chengwei Huang, Zhen Cui, Yuan Zong, Mengying Chen, Qiang Hao, Yi Liu, Jindu Zhu, and

Haibo Li. FENP: A database of neonatal facial expression for pain analysis. *IEEE Transactions on Affective Computing*, 14(1):245–254, 2023.

[126] Guanming Lu, Haoxia Chen, Jinsheng Wei, Xiaonan Li, Xianlan Zheng, Hongyao Leng, Yimo Lou, and Jingjie Yan. Video-based neonatal pain expression recognition with cross-stream attention. *Multimedia Tools and Applications*, 83:4667–4690, 2024.

[127] Denise Harrison, Margaret Sampson, Jessica Reszel, Koowsar Abdulla, Nick Barrowman, Jordi Cumber, Ann Fuller, Claudia Li, Stuart Nicholls, and Catherine M Pound. Too many crying babies: a systematic review of pain management practices during immunizations on youtube. *BMC pediatrics*, 14(1):134, 2014.

[128] Denise Harrison, Shokoufeh Modanloo, Ashley Desrosiers, Louise Poliquin, Mariana Bueno, Jessica Reszel, and Margaret Sampson. A systematic review of youtube videos on pain management during newborn blood tests. *Journal of Neonatal Nursing*, 24(6):325–330, 2018.

[129] Luigi Celona and Luca Manoni. Neonatal facial pain assessment combining hand-crafted and deep features. In *International Conference on Image Analysis and Processing*, pages 197–204. Springer International Publishing, 2017.

[130] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*, pages 1–12. British Machine Vision Association, 2015.

[131] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, page 503–510. Association for Computing Machinery, 2015.

[132] Ghada Zamzmi, Dmitry Goldgof, Rangachar Kasturi, and Yu Sun. Neonatal pain expression recognition using transfer learning. *arXiv preprint arXiv:1807.01631*, 2018. Accessed July 28, 2025.

[133] Ghada Zamzmi, Rahul Paul, Dmitry Goldgof, Rangachar Kasturi, and Yu Sun. Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN). In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[134] 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 (CVPR)*, pages 770–778, June 2016.

[135] Lucas P Carlini, Leonardo A Ferreira, Gabriel AS Coutrin, Victor V Varoto, Tatiany M Heiderich, Rita CX Balda, Marina CM Barros, Ruth Guinsburg, and Carlos E Thomaz. A convolutional neural network-based mobile application to bedside neonatal pain assessment. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 394–401. IEEE, 2021.

[136] Tatiany Marcondes Heiderich, Ana Teresa Figueiredo Stochero Leslie, and Ruth Guinsburg. Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. *Acta Paediatrica*, 104(2):e63–e69, 2015.

[137] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3319–3328. PMLR, 2017.

[138] Mingze Sun, Haoxiang Wang, Wei Yao, and Jiawang Liu. AuE-IPA: An AU engagement based infant pain assessment method. *arXiv preprint arXiv:2212.04764*, 2022. Accessed July 28, 2025.

[139] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2017. Accessed July 28, 2025.

[140] Ghazal Bargshady, Calvin Joseph, Niraj Hirachan, Roland Goecke, and Raul Fernandez Rojas. Acute pain recognition from facial expression videos using vision transformers. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE, 2024.

[141] Manuel Benavent-Lledo, Maria Dolores Lopez-Valle, David Ortiz-Perez, David Mulero-Pérez, Jose Garcia-Rodriguez, and Alexandra Psarrou. Multimodal pain assessment with transformers. *Neurocomputing*, page 132066, 2025.

[142] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 10078–10093. Curran Associates, Inc., 2022.

[143] Melpo Pittara, Andreas Anastasiou, Konstantinos Andreou, Andreas Panayides, Nicolai Petkov, and Constantinos S Pattichis. Facial image and video pain intensity estimation. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pages 937–944. IEEE, 2024.

[144] Paola Lago, Elisabetta Garetti, Giovanna Boccuzzo, Daniele Merazzi, Anna Pirelli, Luisa Pieragostini, Simone Piga, Marina Cuttini, and Gina Ancora. Procedural pain in neonates: the state of the art in the implementation of national guidelines in italy. *Pediatric anesthesia*, 23(5):407–414, 2013.

[145] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556v6*, 2015. Accessed October 27, 2025.

[146] Jacqueline Hausmann, Md Sirajus Salekin, Ghada Zamzmi, Dmitry Goldgof, and Yu Sun. Robust neonatal face detection in real-world clinical settings. *arXiv preprint arXiv:2204.00655*, 2022. Accessed July 28, 2025.

[147] Ethan Grooby, Chiranjibi Sitaula, Soodeh Ahani, Liisa Holsti, Atul Malhotra, Guy A Dumont, and Faezeh Marzbanrad. Neonatal face and facial landmark detection from video recordings. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2023.

[148] TensorFlow. Classification on imbalanced data. https://www.tensorflow.org/tutorials/structured_data/imbalanced_data. Accessed February 23, 2024.

[149] François Chollet. Grad-CAM class activation visualization. https://keras.io/examples/vision/grad_cam/. Accessed February 23, 2024.

[150] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology*, 162:110786, 2023.

[151] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8930–8941. Curran Associates, Inc., 2019.

[152] Pieter C Gort, Cris HB Claessens, Fons van der Sommen, et al. Evaluating the interpretability of prototype networks for medical image analysis. In *Medical Imaging 2025: Image Processing*, volume 13406, page 134061I. SPIE, 2025.

[153] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence*, 5:825565, 2022.

[154] Francesco Del Monte, Roberta Barolo, Maria Circhetta, Angelo Giovanni Delmonaco, Emanuele Castagno, Emanuele Pivetta, Letizia Bergamasco, Matteo Franco, Gabriella Olmo, and Claudia Bondone. Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study. *Frontiers in Digital Health*, 7:1624786, 2025.

[155] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2025. Accessed June 23, 2025.

[156] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. Accessed August 31, 2024.

[157] OpenAI, Inc. ChatGPT. https://chatgpt.com/. Accessed August 31, 2024.

[158] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2025. Accessed August 31, 2024.

[159] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. Accessed August 31, 2024.

[160] Carl Preiksaitis, Nicholas Ashenburg, Gabrielle Bunney, Andrew Chu, Rana Kabeer, Fran Riley, Ryan Ribeira, and Christian Rose. The role of large language models in transforming emergency medicine: scoping review. *JMIR medical informatics*, 12:e53787, 2024.

[161] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. *arXiv preprint arXiv:2401.06775*, 2024. Accessed August 31, 2024.

[162] Kaan Can Demirbaş, Mehmet Yıldız, Seha Saygılı, Nur Canpolat, and Özgür Kasapçopur. Artificial intelligence in pediatrics: learning to walk together. *Turkish Archives of Pediatrics*, 59(2):121, 2024.

[163] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1), 2023.

[164] Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of internal medicine*, 177(2):210–220, 2024.

[165] Zahir Kanjee, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80, 2023.

[166] Takanobu Hirosawa, Yukinori Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura, and Taro Shimizu. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *International journal of environmental research and public health*, 20(4):3378, 2023.

[167] Takanobu Hirosawa, Ren Kawamura, Yukinori Harada, Kazuya Mizuta, Kazuki Tokumasu, Yuki Kaji, Tomoharu Suzuki, and Taro Shimizu. ChatGPT-generated differential diagnosis lists for complex case–derived clinical vignettes: Diagnostic accuracy evaluation. *JMIR Medical Informatics*, 11:e48808, 2023.

[168] Takanobu Hirosawa, Yukinori Harada, Kazuya Mizuta, Tetsu Sakamoto, Kazuki Tokumasu, and Taro Shimizu. Diagnostic performance of generative artificial intelligences for a series of complex case reports. *Digital Health*, 10:20552076241265215, 2024.

[169] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, Inc., 2017.

[170] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

[171] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Preprint, OpenAI, 2018. Accessed October 27, 2025.

[172] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[173] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901. Curran Associates, Inc., 2020.

[174] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. Accessed September 4, 2025.

[175] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2022. Accessed September 4, 2025.

[176] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 27730–27744. Curran Associates, Inc., 2022.

[177] OpenAI, Inc. Introducing ChatGPT. https://openai.com/index/chatgpt/, 2022. Accessed September 4, 2025.

[178] OpenAI, Inc. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed August 31, 2024.

[179] OpenAI, Inc. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed August 31, 2024.

[180] OpenAI, Inc. Introducing GPT-4.5. https://openai.com/index/introducing-gpt-4-5/, 2025. Accessed September 4, 2025.

[181] OpenAI, Inc. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed September 4, 2025.

[182] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. Accessed September 4, 2025.

[183] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. Accessed September 4, 2025.

[184] Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. Accessed June 23, 2025.

[185] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. Accessed September 4, 2025.

[186] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. Accessed September 4, 2025.

[187] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. Accessed June 23, 2025.

[188] Meta AI. Llama 4: Advancing multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed September 4, 2025.

[189] ArtificialAnalysis. Comparison of models: Intelligence, performance & price analysis. https://artificialanalysis.ai/models. Accessed August 31, 2024.

[190] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[191] Google LLC. Gemini. https://gemini.google.com/app. Accessed August 31, 2024.

[192] Google LLC. Google AI Studio. https://aistudio.google.com/. Accessed August 31, 2024.

[193] Meta Platforms, Inc. LLaMA 3 (8B) on Ollama. https://ollama.com/library/llama3:8b. Accessed August 31, 2024.

[194] Emre Sezgin. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. *Digital health*, 9:20552076231186520, 2023.

[195] Lisa Clayton, Mike Wells, Scott Alter, Joshua Solano, Patrick Hughes, and Richard Shih. Educational concepts: A longitudinal interleaved curriculum for emergency medicine residency training. *JACEP Open*, 5(3):e13223, 2024.

[196] Joseph Barile, Alex Margolis, Grace Cason, Rachel Kim, Saia Kalash, Alexis Tchaconas, and Ruth Milanaik. Diagnostic accuracy of a large language model in pediatric case studies. *JAMA pediatrics*, 178(3):313–315, 2024.

[197] Sarah KS Knack, Nathaniel Scott, Brian E Driver, Matthew E Prekker, Lauren Page Black, Charlotte Hopson, Ellen Maruggi, Olivia Kaus, Walker Tordsen, and Michael A Puskarich. Early physician gestalt versus usual screening tools for the prediction of sepsis in critically ill emergency patients. *Annals of Emergency Medicine*, 84(3):246–258, 2024.