POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Pain assessment in infants via facial expressions analysis and deep learning

Supervisors

Prof. Gabriella OLMO

Letizia BERGAMASCO

Candidate

Marta LATTANZI

Academic Year 2023/24

Abstract

This research presents a deep learning-based approach for infant pain assessment and it was conducted in LINKS Foundation in collaboration with the Neonatal Unit of AO Ordine Mauriziano Hospital and the Pediatric Emergency Department of Regina Margherita Hospital, all based in Turin. Since young children are unable to communicate verbally the experience of pain, accurate pain assessment using validated tools is crucial to determine the most effective pain management strategies. Traditional pain assessment relies on the use of pain scales that consider behavioural, physiological, and contextual indicators, but none of these scales has yet emerged as the gold standard. The use of these tools is also limited by the lack of objectivity and repeatability of the assessment, which depends on the experience of healthcare professionals and the simultaneous monitoring of numerous parameters. For these reasons, automated pain assessment systems are highly desirable in clinical practice for more efficient recognition and management of pain. The aim of this thesis is to propose an objective and contactless computer vision approach for infant pain evaluation based on facial expression analysis. Video data related to two distinct infant populations are used: newborns undergoing heel stick procedures for blood sampling, and children aged 3-36 months admitted to the Pediatric Emergency Department with acute pain. Videos are accompanied by the corresponding pain scores assigned by healthcare professionals using traditional pain scales. Two separate datasets are created by extracting frames from the video recordings and labelling each frame based on the assigned score. Each dataset is used to train a specific convolutional neural network for binary pain classification. For both infant populations, the trained models achieve high accuracy in classifying images into pain or nonpain categories. Moreover, the application of a visual explanation technique shows that the networks' decisions are based on the facial regions most closely associated to the experience of pain. In conclusion, our promising findings pave the way for the development of an automated system that integrates, standardises, and improves human pain evaluation.

Acknowledgements

I would like to thank Professor Gabriella Olmo for her availability and valuable advice and Letizia Bergamasco for her support and guidance throughout this thesis. I am very grateful to have had the opportunity to work with them on this project. I would also like to thank the LINKS Foundation for welcoming me during these months.

Table of Contents

List of Tables iv

List of Figures v

1 Introduction 1
- 1.1 Thesis Purpose . . . 1
- 1.2 Pain Assessment . . . 1
  - 1.2.1 Physiology of Pain . . . 1
  - 1.2.2 Pain in Infants . . . 2
  - 1.2.3 Traditional Methods for Evaluation of Pain in Infants . . . 4
  - 1.2.4 Importance of Automatic Infant Pain Assessment based on Facial Expression . . . 6
- 1.3 Thesis Overview . . . 7

2 Literature Review 8
- 2.1 Infant Facial Expressions of Pain . . . 8
- 2.2 Existing Datasets with Pain Score Annotations . . . 10
- 2.3 Automatic Pain Classification in Infants . . . 15
  - 2.3.1 Related Works . . . 16

3 Deep Learning Theoretical Framework 19
- 3.1 Overview . . . 19
- 3.2 Deep Learning . . . 20
  - 3.2.1 Artificial Neural Networks . . . 20
- 3.3 Convolutional Neural Networks . . . 22
  - 3.3.1 VGG-16 . . . 25
- 3.4 Explainability in Deep Learning . . . 27
  - 3.4.1 Grad-CAM . . . 27

4 Materials and Methods 30
- 4.1 Data Description and Pain Scale Implementation . . . 30

4.1.1 M-PAIN Dataset 30
4.1.2 VideoDol Dataset 32
4.2 Comparative Study of Face Detectors 34
4.2.1 Face Detection 34
4.2.2 State-Of-The-Art Face Detectors 34
4.2.3 Performance Comparison of Face Detectors 36
4.3 Pain Classification 37
4.3.1 Computational Resources and Framework 38
4.3.2 CNN Model 38
4.3.3 CNN for Pain Classification in Newborns 42
4.3.4 CNN for Pain Classification in Pediatric ED 45
4.4 Explainability 46

5 Results and Discussion 48
5.1 CNN for Pain Classification in Newborns 48
5.2 CNN for Pain Classification in Pediatric ED 52
5.3 Results Discussion and Limitations 54

6 Conclusion and Future Works 56

A Evaluations 58

Bibliography 60

List of Tables

1.1 DAN Scale. P: period of observation. Source: [21] 5
1.2 FLACC scale. Source: [19] 6

2.1 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. 14

4.1 Results of comparative study of face detectors 37
4.2 Confusion matrix illustrating the model predictions against true class labels 41

5.1 Cross-validation results for *pain* vs *nonpain* classification, in terms of accuracy and F1-score (mean ± standard deviation). 48

A.1 FLACC face score for VideoDol dataset 59

List of Figures

1.1 Pain pathway. Source [2] 2

2.1 Primal face of pain, involving brow bulge, eye squeeze, and a horizontally stretched open mouth with deepening of the nasolabial furrow. Source: [32] 9
2.2 Examples of iCOPE images labelled as nonpain 11
2.3 Examples of iCOPE images labelled as pain 11

3.1 Relationship between AI, ML and DL 20
3.2 Neural Network Architecture 21
3.3 Example of convolutional operation 23
3.4 Example of pooling operations 25
3.5 VGG-16 model architecture 26
3.6 Grad-CAM examples. Source: [59] 28

4.1 Example image of M-PAIN dataset during the blood sampling procedure. The image was postprocessed to anonymise infant face. . . 32
4.2 Example image of VideoDol dataset setup. The image was postprocessed to anonymise infant face. 33
4.3 CNN model architecture 39
4.4 Overview of the complete training CNN process 40
4.5 StratifiedGroupKFold on M-PAIN dataset (training set includes both training and validation data). 44
4.6 Architecture to describe the Grad-CAM technique for generating the heatmap from an input image. Adapted from [59] 47

5.1 Examples of resulting Grad-CAM heatmaps on M-PAIN test set images. Labels are reported below each image. The images were postprocessed to anonymise infant faces. 50
5.2 Examples of resulting Grad-CAM heatmaps on M-PAIN test set images of newborns without pacifier. Labels are reported below each image. The images were postprocessed to anonymise infant faces. . 50

5.3 Examples of resulting Grad-CAM heatmaps on iCOPE test set. Labels are reported below each image. 51
5.4 Examples of resulting Grad-CAM heatmaps on M-PAIN test set images with training cross-database. Labels are reported below each image. The images were postprocessed to anonymise infant faces. . 52
5.5 Normalized confusion matrix of VideoDol test set. 53
5.6 Examples of resulting Grad-CAM heatmaps on VideoDol test set images correctly classified. Labels are reported below each image. The images were postprocessed to anonymise infant faces. 53
5.7 Examples of resulting Grad-CAM heatmaps on misclassified test set images of VideoDol dataset. Labels passed to Grad-CAM are reported below each image. The images were postprocessed to anonymise infant faces. 54

Chapter 1

Introduction

1.1 Thesis Purpose

This thesis investigates the development of deep learning techniques to automatically assess pain in newborns and young children. The aim is to enhance the objectivity and convenience of pain assessment with a contactless computer vision approach that analyses infant facial expressions.

1.2 Pain Assessment

1.2.1 Physiology of Pain

Pain is more than just an unpleasant sensation, it is a complex sensory system that plays a crucial role in our interaction with the external environment, allowing us to avoid potential dangers and injuries. When the nervous system detects stimuli capable of damaging tissues, it triggers immediate reflex reactions. Between the damaging stimulus at the tissue level and the subjective experience of pain various chemical and electrical events occur, which can be divided into four phases: transduction, transmission, modulation and perception [1], as shown in Figure 1.1. In the first phase the nerve endings called nociceptors, which are present in nearly every tissue, respond to various forms of energy, activating when the stimulus intensity exceeds a certain threshold. In the transmission phase the information detected by the nociceptors is transported in the form of bio-electric signals to the spinal cord and then reach the cerebral cortex. Modulation refers to the control of pain transmission by neurological activity, influencing the intensity and quality of pain perception. Finally, perception is the subjective experience of pain, involving various brain structures and emotional factors that can impact the perception of pain [1].

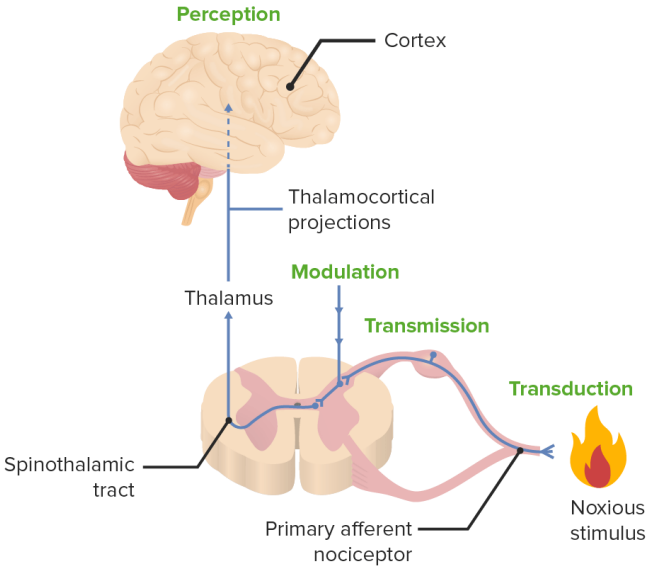


Figure 1.1: Pain pathway. Source [2]

Consequently, pain is not reducible to the simple conduction of the stimulus but is the result of a complex interaction between different structures and phenomena that continually modulate the amplitude and quality of perception [1]. Understanding the complexity of pain perception is crucial, not only for managing its impact on individuals, but also for ensuring its proper assessment and treatment in healthcare settings. This is underscored by the Italian Law 38/2010 [3], which mandates that the characteristics of the pain detected and its development during hospitalisation, as well as the antalgic technique and drugs used, their dosages and the result achieved, must be recorded in each medical record. Hence, pain should be recorded as the fifth vital parameter along with heart rate, respiratory rate, blood pressure and body temperature.

1.2.2 Pain in Infants

In 1979 the International Association for the Study of Pain (IASP) defined the physical pain as "An unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage" [4]. This definition emphasises the "bipolar" nature of pain, consisting of both physiological and psychological variables. However, it aligns more closely with adult experiences, focusing on emotional and sensory aspects that are difficult to evaluate in non-verbal infants. This raises the question of whether such a definition adequately captures the pain experience in early childhood. To address this concern, the IASP Committee on Taxonomy later added this note: "The inability to communicate verbally

does not negate the possibility that an individual is experiencing pain and is in need of appropriate pain-relieving treatment" [5]. Only recently, the IASP introduced a revised definition of pain as "An unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" [6].

However, until the 1980s, researchers largely believed that newborn babies could not feel pain due to the perceived immaturity of their central and peripheral nervous systems [7]. This outdated belief have heavily conditioned the approach to children with painful symptoms, both in hospitals and at home. As a result, pain in infants, especially in pre-term children, has often been underdiagnosed and consequently undertreated [8]. On the other hand, recent research indicates that newborns, including premature ones, feel pain and remember pain experiences [9]. In fact, current scientific understanding indicates that, by the 24th week of gestation, a fetus possesses all the anatomical and neurochemical features required for pain perception. The nociceptive system further develops after birth, reaching maturity around the first year of life. The antinociceptive system, responsible for pain modulation, matures more slowly. Therefore, in response to an equivalent painful stimulus, younger patients experience a greater perception of pain due to reduced central and peripheral inhibition [1]. Additionally, studies have revealed that repeated exposure to pain during early nervous system development can result in various negative effects on physiology, cognition, behavior, hormones, and the endocrine system, both in the short and long term [10] [11] [12]. It follows that untreated pain exposes the infant to a high risk of complications, so the assessment of pain is crucial, especially in order to identify possible interventions for its prevention and treatment.

Pain in infants can be categorized into three main types: acute procedural, acute prolonged (including post-operative pain) and chronic pain [13]. Acute procedural pain refers to short-term pain caused by a painful stimulus, such as immunization, that quickly dissipates. Acute prolonged pain follows a surgical procedure and continues for a longer period. In contrast, chronic pain is a recurring pain experience that is not associated with a specific event and has a significantly longer duration than prolonged acute pain. Non-pharmacological analgesia techniques before and during the performance of a painful procedure have proven to be effective in containing the newborn's pain, making it possible to limit the use of analgesic drugs. These techniques include containment facilitated through the use of the operators' hands (holding) or through the use of sheets or blankets (wrapping) that limit postural instability and attenuate physiological and behavioural responses to pain. The administration of sweetened substances (such as sucrose, glucose or breast milk) and the skin-to-skin contact are also effective in reducing the painful sensation [14].

1.2.3 Traditional Methods for Evaluation of Pain in Infants

Due to the inability of the newborn to express pain verbally, the quantitative assessment of neonatal pain is extremely difficult and relies mainly on the evaluation of multiple parameters. Over the years, several pain assessment scales have been validated to diagnose, quantify and monitor neonatal pain across different newborn populations (term, preterm), as well as for different types of pain such as procedural pain and postoperative pain.

The algometric scales can be divided into two types: unidimensional scales, which focus on either behavioural or physiological parameters, and multidimensional scales, which consider both behavioural and physiological aspects. Behavioural parameters include facial expression, limb movement, and vocalization, while physiological measurements include heart rate, respiratory rate, oxygen saturation, and blood pressure. The multidimensional approach enhances the sensitivity and specificity of the scales, contributing to a more accurate understanding of the pain experience [15]. However, the current trend is to attribute less importance to physiological indicators in favour of behavioural ones and, in particular, facial expressions, which are more accurate in distinguishing painful situations from non-painful. Behavioral indicators appear to be more significant because, from an evolutionary perspective, they are specifically designed to capture the caregiver's attention [15]. The facial expressions include brow bulge, eye squeeze, nasolabial furrow and a vertically stretch mouth with a lowered jaw. Widely used monodimensional scales include the Neonatal Facial Coding System (NFCS) [16] [17], the Echelle Douleur Inconfort Nouveau-Né (EDIN) [18], and the Face, Legs, Activity, Cry, Consolability (FLACC) scale [19]. Among multidimensional scales, frequently used ones are the Premature Infant Pain Profile (PIPP) [20], Douleur Aiguë du Nouveau-né (DAN) [21] and Neonatal Infant Pain Scale (NIPS) [22]. The next paragraphs focus specifically on two algometric scales: the DAN scale and the FLACC scale as they are the pain assessment tools used by healthcare professionals in the context of this study.

DAN Scale

The DAN scale (Table 1.1), developed in 1996, was validated through 42 venous or heel stick samples for glucose measurement in newborns with respiratory autonomy [21]. This scale allows assigning a score based on three factors: facial response, limb movements, and vocal expression, each scored from 0 to 4, 0 to 3, and 0 to 3, respectively, as represented in Table 1.1. Thus, the total score ranges from 0 to 10, a rating greater than or equal to 3 is associated with an experience of pain. This scale does not require instrumentation and relies solely on clinical observation in the 30 seconds following the painful procedure. Specifically focusing on the assessment of facial expressions, if the newborn is calm, a score of 0 is assigned,

while if the newborn whines, a score of 1 is assigned. For higher scores, three actions out of nine in NFCS (Neonatal Facial Coding System) [16] [17] are considered: eye squeezing, brow bulging, and accentuation of the nasolabial furrow. Depending on the intensity of one or more of these parameters, the painful expression can be classified as mild (score 2), moderate (score 3), or severe (score 4) [21].

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, and 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	

Table 1.1: DAN Scale. P: period of observation. Source: [21]

FLACC Scale

The FLACC scale [19] (Table 1.2) is a combination of 5 behaviors: Face, Legs, Activity, Cry and Consolability, that are considered indicative of pain and can be detected and graded by an observer. Each behavior is assigned a score on a scale of 0 to 2, leading to a pain intensity score ranging from 0 to 10. The initial usage guidelines advised observing the child for 1 to 5 minutes and correlating the observed behaviors with those outlined in the scale for each item. Despite

the Ministry of Health [1] recommending its use in assessing pain in individuals under the age of 3, including the neonatal period, its validation was conducted on a sample of 89 children aged between 2 months and 7 years undergoing various surgical procedures [19].

Category	**Scoring**		
	0	**1**	**2**
Face	No particular expression or smile	Occasional grimace	Frequent to 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 (awake or asleep)	Moans or whimpers; occasional complaint	Crying steadily, screams, sobs, frequent complaints
Consolability	Content, relaxed	Reassured by touching, hugging or being talked to, distractible	Difficult to console or comfort

Table 1.2: FLACC scale. Source: [19]

1.2.4 Importance of Automatic Infant Pain Assessment based on Facial Expression

Despite the usefulness of the traditional algometric scales, their actual application in clinical practice is limited. No scale has been recognised as the gold standard, and there are no clear indications on which is the best to use [15]. The use of pain scales is also limited by the poor objectivity and repeatability of the assessment, which are strictly dependent on the sensitivity and experience of the healthcare professionals, complicated by the simultaneously monitoring of multiple parameters. A recent survey of Italian emergency departments [23] revealed positive progress in pain measurement, with increased utilisation of algometric scales. However, the study also indicates that achieving optimal pain management remains a challenge

in these settings, necessitating further efforts for improvement. Therefore, the implementation of new methods with easy clinical applicability that allow for an objective and repeatable assessment of pain in infants is highly desirable. A recent review [24] highlights the importance of developing automated face-based pain detection systems for infants, emphasizing the gap between computational methods developed for automated pain assessment and their real-time bedside application.

Facial expressions are one of the most important indicators of pain in people with limited verbal communication, such as infants. Moreover, in the case of newborns undergoing painful procedures, containment methods (wrapping and use of pacifier) are typically employed, making it challenging to analyse other factors, such as vocal expression or body movements. By concentrating exclusively on facial expressions, automatic pain detection becomes feasible, aligning with containment methods while remaining contactless and non-intrusive, in order to avoid potential complications linked to sensors directly placed on infants' skin, such as skin damage or an elevated risk of infection transmission.

Starting from previous works of [25] [26], this thesis aims to propose an approach to assess pain in infants using facial expression analysis and deep learning techniques. This approach will be fundamental to developing a more convenient and objective system for automatic pain assessment in infants.

The project has been developed in LINKS Foundation, a research center based in Turin, in collaboration with the Neonatal Unit of AO Ordine Mauriziano Hospital and the Pediatric Emergency Department of Regina Margherita Hospital. Every decision made in this work has been agreed upon with healthcare professionals to ensure a multidisciplinary view of the problem.

1.3 Thesis Overview

The following Chapter describes existing datasets with pain score annotations and various techniques used in the literature for automatic pain assessment, including a Section outlining the state of the art. Chapter 3 provides a theoretical framework for the thesis, with a specific focus on deep learning, convolutional neural networks, and explainable artificial intelligence methods. Chapter 4 describes two datasets under study in this thesis, including infant video recordings with expert pain annotations and a detailed explanation of the proposed approach for binary pain classification using deep learning models. Later in Chapter 5, the experimental results are presented and discussed. Finally, Chapter 6 presents the conclusions of the thesis and identifies potential avenues for further improvement.

Chapter 2

Literature Review

This Chapter examines facial pain expressions in infants, first exploring their nature and importance in assessing pain. It then reviews existing neonatal datasets annotated with pain scores and various techniques used in the literature for automated pain assessment in infants, providing a comprehensive analysis of the most relevant methods and approaches.

2.1 Infant Facial Expressions of Pain

Among the non-verbal ways to communicate emotional state, it is important to consider face and gaze. Indeed, facial expressions can be considered a spontaneous, immediate and involuntary response to the emotions felt by the subject. According to [27], facial expression are a primary and perhaps the most consistent cue when assessing pain in children, even above cry. Charles Darwin was the first to use a scientific approach to study the correlation between facial movements and emotional experience [27], but it was not until the 1970s that the main techniques for detecting and analysing facial expressions emerged.

In 1977, Ekman and Frisen, based on the measurement studies of the actions of the facial musculature of Hjortsjo [28], developed the Facial Action Coding System (FACS) [29]. In addition, Ekman was the first to propose that these expressions are innate and reflexive, meaning that even neonates can use them involuntarily to capture attention. FACS is a system that describes the movement of facial muscles and how each muscle acts in visibly changing the configuration of the face. Each facial movement is the result of the action of an exact number of muscles that contract or relax synergistically. Each individual detectable movement, given by the contraction or relaxation of one or more muscles, is described by Action Units (AUs), which make it possible to encode any facial expression and analyse the

different emotional states. The coding manual explains specifically interpretation and meaning of each individual AU and also classifies a total of 44 movements [27].

The challenges associated with the use of the FACS has led to the development of other coding systems tailored specifically for examining pain in children [27]. These facial coding systems are:

- MAX (Maximally Discriminative Facial Movement Coding System), that analyses three different anatomical areas (forehead and eyebrows, eyes and nose, and mouth and chin) to estimate various emotions, including pain [30].
- NFCS (Neonatal Facial Coding System), that includes 10 facial actions associated with pain and derived from the FACS [16].
- CFCS (Child Facial Coding System), that focuses on 13 facial actions and it derived from both the FACS and NFCS to identify pain in toddlers and school-age children [31].

In 2003, Oster [33] proposed a FACS adaptation for infants (Baby FACS), that identifies anatomical variation in the facial structures of children. Baby FACS includes an additional action unit in the brow area, along with adjustments guided by the difference in facial morphology and dynamics between adults and infants. Later, in 2008, Schiavenato [27] hypothesised that newborns have a Primal Face of Pain (PFP), an instinctive, innate and universal expression that allows the child to communicate stressful situations such as pain. This expression include mouth opening, brows furrowing and eyes closing. The PFP serves as the foundation for recognising pain in newborns, unaffected by social factors that may influence painful

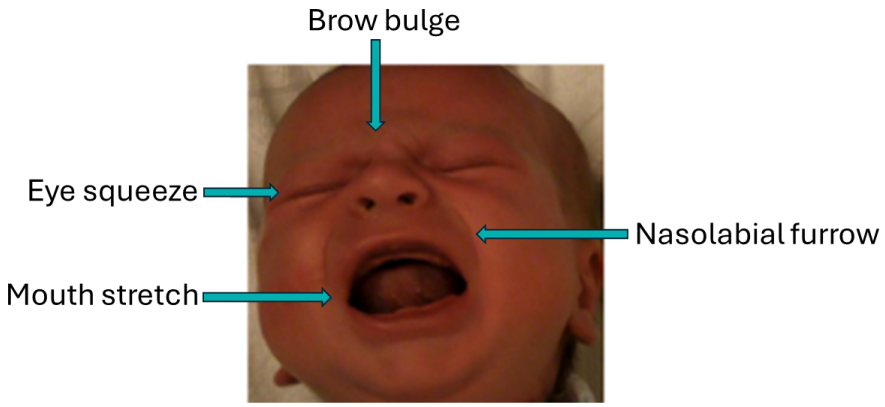


Figure 2.1: Primal face of pain, involving brow bulge, eye squeeze, and a horizontally stretched open mouth with deepening of the nasolabial furrow. Source: [32]

expressions through the control and alteration of specific movements. Newborns faithfully display the PFP to communicate pain, whereas adults tend to modify expressions based on their experiences and the environment in which they were raised. Thus, Schiavenato highlights that it is no longer suitable to reference the PFP in older children aged over 6 [27].

Despite the existence of different pain coding models of facial expressions, they often evaluate highly similar, and frequently identical, parameters. This phenomenon arises from the universal nature of emotional communication, as humans possess an innate understanding of micro facial expressions. The infant pain expression (Figure 2.1) is always associated with eye squeeze, brow bulge, and a horizontally stretched open mouth with deepening of the nasolabial furrow [27].

2.2 Existing Datasets with Pain Score Annotations

According to a recent review [24], there is a limited number of datasets for facial expression analysis of pain in newborns. These databases are extensively used in the academic scientific environment; however, they have certain limitations. These include a restricted quantity of images, a representation of a single ethnic group, limited confidence due to insufficient explanations regarding ethics committee approval, and a predominant focus on a specific clinical population, predominantly term newborns, not allowing studies with critically ill and preterm newborns [24]. Several databases have a limited number of newborns, limited types of pain stimuli that differ greatly from painful clinical procedures [34] or they have not collected the data under different levels of pain [35]. Despite their limitations, these datasets offer valuable contributions to the field of research. This Section provides brief descriptions of the publicly available datasets specifically focused on neonatal pain, with a summarised overview presented in Table 2.1.

iCOPE

The Infant COPE (Classification Of Pain Expressions) database [36] [37] [38] [39] contains a total of 204 images of 26 Caucasian neonates (13 boys and 13 girls) ranging in age from 18 hours to 3 days. The facial expressions of the infants were photographed in one session while the neonates were experiencing four distinct stimuli in the following sequence:

- Rest/cry: after moving the neonates to a new crib, they were swaddled and their behavior (crying or resting) was documented at the time of each picture.

- Air stimulus: a small amount of air from a squeezable plastic camera lens cleaner was directed towards the newborns' nose.
- Friction: a cotton wool soaked in alcohol was rubbed on the outer side of the newborns' heel.
- Pain: the external lateral surface of the newborns' heel was punctured for blood collection.

The photographs taken after the series of the first three stress-inducing stimuli were classified as *nonpain*, while the photographs corresponding to the heel puncture was classified as *pain*. The iCOPE dataset presents a unique challenge due to the presence of crying images caused by both pain and non-painful stimuli, such as discomfort or frustration. Figures 2.2, 2.3 show some example images of iCOPE dataset.

(a) Rest **(b)** Air **(c)** Cry **(d)** Friction **(e)** Friction

Figure 2.2: Examples of iCOPE images labelled as nonpain

(a) Pain **(b)** Pain

Figure 2.3: Examples of iCOPE images labelled as pain

iCOPEVid

The Infant Cope Video dataset (iCOPEVid) [40] expands the data collection of iCOPE and it contains 234 videos (20 seconds/video) obtained following the same procedures. Video images were collected from 49 newborns (26 boys and 23 girls) aged between 34 to 70 hours. Ethnicity was distributed as 73 Caucasian, 5 African and 1 Asian.

NPAD

Zamzmi et al. [41] collected data of 40 neonates (approximately 50 percent male and 50 percent female), aged between 32 and 40 weeks of gestational age (GA), within the NICU environment. They have different ethnicity: 13 White, 8 Caucasian, 6 African-American and 4 Asian. Among them, 31 neonates were recorded undergoing procedural pain (routine heel lancing and immunization), while the remaining 9 undergoing postoperative pain, for example, due to a gastronomy tube procedure. In addition to video recordings capturing the neonate's face, head, and body, as well as audio signals, vital sign readings were collected. For procedural pain the score are assigned based on the NIPS scale [22], categorising pain into three labels: no pain (score 0-2), moderate pain (3-4), and severe pain (>4). While for the post operative pain, the N-PASS scale [42] was used, providing five emotional states, that are deep sedation, light sedation, normal, mild pain, and severe pain. This dataset is highly representative of the real-world condition as it was obtained during standard clinical procedures in a typical clinical setting.

APN-db

The Acute Pain in Neonates (APN-db) database [43] contains more than 200 videos of infants undergoing different painful and painless procedures. It is divided in two parts: the vaccination pain partition and the NICU partition. In the first partition, 112 infants, aged between 0 and 6 months, were recorded during different types of vaccines via intramuscular injection. While in the Neonatal Intensive Care Unit partition, a total of 146 procedures involving 101 individual participants (GA 26-41) were collected. The videos were annotated frame by frame with the Neonatal Face and Limb Acute Pain Scale (NFLAPS), which integrates facial expression parameters from the NFCS scale [16] [17] with limb movement parameters (arm and leg movements) from the NIPS scale [22]. This scale assesses pain intensity on a scale ranging from 0 to 11.

FENP

The facial expression of neonatal pain (FENP) database [35] is composed by 11000 images of 106 different Chinese neonates, aged between 2 days and 4 weeks. The facial expression images encompass four distinct categories: calmness expression, crying expression, mild pain expression and severe pain expression, each comprising 2750 neonatal facial expression images. For the pain emotion category, videos were collected when neonates underwent intramuscular injections or blood sampling on their heels. The crying emotion category includes scenes of changing the baby crib and other crying situations not induced by pain (such as hunger or fear). The calmness category comprises videos of sleeping or calm neonates. To classify the

neonatal pain images into two pain degrees (mild pain or severe pain), the NFCS scale [16] [17] was used and images with scores between 6 and 10 were categorized as severe pain, while those with scores between 1 and 5 were labeled as mild pain.

DFEPN

The Dynamic Facial Expression of Pain in Neonates (DFEPN) dataset [44] consists of 1897 video clips with four different categories of expression: calmness, crying, moderate pain, and severe pain. These clips were recorded from 106 neonates using the same procedure and labelling method as the FENP dataset [35]. The length of each video ranges from 20 to 60 seconds.

USF-MNPAD-I

The USF-MNPAD-I (University of South Florida Multimodal Neonatal Pain Assessment Dataset) [45] includes recordings from 58 neonates (GA 27-41) during hospitalization in the neonatal intensive care unit. It contains video, audio signals, physiological signals and also the contextual data. The dataset includes individuals from diverse ethnic and racial backgrounds, with an almost equal representation of males and females. It covers both procedural and postoperative pain scenarios, with data from 36 subjects undergoing 44 procedures and 9 subjects in the postoperative phase. Additionally, cortical activity data using NIRS were collected from a subset of neonates. Pain assessments were conducted using the NIPS scale [22] for procedural pain (scored every minute) and the N-PASS scale [42] for postoperative pain.

YouTube Immunization and YouTube Blood Test

The YouTube Immunization dataset [46] is composed by 142 YouTube videos, filmed by parents or guardians, capturing infants (0-12 months) undergoing immunization injections. Expert evaluators utilised the FLACC pain scale [19] to assess the videos at three specific time points: 15 seconds before the initial injection, during the initial and the final injection, with an additional assessment 15 seconds after the last injection. Supplementary data such as the infant's gender, number of injections, and caregiver's gender were also collected.
Using a similar approach, in 2018, Harrison et al. [47] collected YouTube videos of babies undergoing blood test. A total of 55 videos, showing 63 procedures, were selected for analysis, with evaluations conducted using Neonatal Facial Coding System (NFCS) scores [16] [17].
The main limitation of these databases is the poor quality of the captured videos, which results in many videos being excluded from annotation. In addition, some

videos are no longer available (offline or non-public) and in other videos the baby's face is not visible at certain times points.

Dataset	**Subjects**	**Age**	**Ethnicity**	**Size**	**Type**	**Pain labels**
iCOPE [36] [37] [38] [39]	26	18-36 h	Caucasian	204 images	PR	pain/nonpain
iCOPEvid [40]	49	34-70 h	Diversified	234 videos	PR	pain/nonpain
NPAD [41]	40 (PR+PO)	GA: 32-40 w	Diversified	multi-modal data	PR, PO	PR: no/moderate/ severe pain, PO: 5 states
USF-MNPAD-I [45]	58 (PR+PO)	GA: 27-41 w	Diversified	multi-modal data	PR, PO	PR: no/moderate/ severe pain, PO: several levels
APN-db [43]	213	GA: 26-41 w 0-26 w	N/A	>200 videos	PR	0-11 pain levels
FENP [35]	106	2 d - 4 w	Chinese	11000 images	PR	calmness/crying/ moderate pain/ severe pain
DFENP [44]	106	N/A	N/A	1897 videos	PR	calmness/crying/ moderate pain/ severe pain
YouTube Imm. [46]	142	0-12 m	N/A	142 videos	PR	0-10 FLACC scores
Youtube Blood Test [47]	63 scenarios	N/A	N/A	55 videos	PR	0-4 NFCS scores

Table 2.1: 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.

2.3 Automatic Pain Classification in Infants

In the last years, different computational methods have been developed to automatically detect the painful phenomenon, to assist healthcare professionals in monitoring pain and identify the need for therapeutic interventions. Current automatic methods for assessing neonatal pain rely on the analysis of facial expressions, crying sounds, body movements, and physiological responses. These approaches can be unimodal, focusing on a single signal modality, or multimodal, integrating multiple modalities for a more accurate and comprehensive assessment of neonatal pain.

Due to their significant importance in communicating pain in newborns, facial expressions have attracted substantial research interest, leading to the development of diverse automatic methods for their analysis. Methods for assessing neonatal pain using facial expressions can be categorised into static and dynamic approaches. Static methods involve extracting relevant features from static images, which are then utilised to train off-the-shelf machine learning classifiers. Existing static approaches can be broadly divided into two groups: handcrafted methods and deep learning-based methods. The former are manually designed by human experts to isolate specific predefined characteristics, they are generally interpretable and relatively efficient but may require design expertise for optimal performance. Two well-known handcrafted methods are: Local Binary Pattern (LBP), which generates a binary pattern by comparing each pixel with its neighbours to form a texture descriptor, and Histogram of Oriented Gradients (HOG), a shape descriptor that computes gradients' intensity and direction within small image regions, thereby constructing a histogram to encapsulate edge details. The challenges and limitations of handcrafted methods have significantly driven the development of deep learning-based methods, in particular Convolutional Neural Networks (CNNs). CNNs autonomously learn and extract pertinent features from raw data or images at various levels of complexity. These networks have demonstrated groundbreaking performance across numerous domains, such as clinical and emotion recognition applications [48].

Other studies [49] [50] have instead used the FACS method for pain assessment. However, it is worth noting that the detection of Action Units (AUs) in infants is very challenging, because the tools for AU detection, trained primarily on adult data, do not performing well on infants, especially in presence of occlusions, like the use of pacifier. Moreover, there are no publicly available databases with AU labels for children and the manual labelling is time-consuming. These challenges complicate the use of the FACS method for assessing pain in infants and may require further development or adaptation to be effective.

The following paragraph provides an overview of literature studies focused on techniques for automatic pain assessment through the analysis of facial expressions in infants.

2.3.1 Related Works

In 2006, pioneering studies [36] [37] [38] were conducted to classify pain based on facial expression. The first paper [36] used three different face classification techniques: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM). The SVM outperformed the other methods, reaching an accuracy of 88% in pain versus non-pain classification on iCOPE dataset using 10-fold cross-validation. However, this work considered a best case scenario where samples of the same neonate were used in both training and testing sets. The above-discussed work is extended in [38] to include NNSOA (Neural Network Simultaneous Optimization Algorithm) for classification along with LDA, PCA, and SVM. Moreover, they used a different evaluation protocol, leave-one-subject-out cross-validation (LOSOXV), that is challenging but more realistic in clinical settings. The results showed that NNSOA achieved the highest average classification rate (90.2%) in classifying infants' images. However, Celona et al. [51] claim that they have reimplemented the methodology of [38] for comparison purposes, reaching an average accuracy of 78.94%.

Instead of detecting the presence or absence of the pain expression, Gholami et al. [52] suggested a method to estimate the intensity level of the pain expression, using a sparse kernel machine algorithm, known as Relevance Vector Machine (RVM). This algorithm, which is a Bayesian adaptation of the Support Vector Machine (SVM), provides the posterior probabilities for class memberships and enabled the generation of pain intensities, that were also compared with the assessment of five expert and five non-expert examiners. For the model's evaluation leave-one-image-out method was applied. The linear kernel RVM algorithm achieved a classification accuracy of 91% on iCOPE dataset, but it failed to converge in 5 of the 21 subjects considered. It is worth noting that Gholami et al. tried to infer intensity of pain on the iCOPE dataset with RVM prior probabilities. The images in the dataset were labeled for pain intensity using a scale from 0 to 100, which does not correlate with any established clinical pain measurement scale for newborns [43].

In 2015, Heiderich et al. [53] proposed a computational framework that enables pain evaluation based on image recognition of pain-related facial actions. The software is capable of identifying 66 facial landmarks and compare the distances between the main nodal points in order to perform facial action detection based on the Neonatal Facial Coding System (NFCS) [16] [17]. If three or more facial actions

were detected, the newborn was considered to be in pain. The results showed 100% of sensitivity and specificity when assessing a neonate with pain and 85% of sensitivity and 100% specificity when assessing a neonate during periods of rest.

Later, Zamzmi et al. [54] implemented a multimodal machine learning based approach that combining facial expression, body movement and changes in vital signs in order to classify the infant's final state as no pain, moderate pain, or severe pain. The method was evaluated using a dataset of 18 neonates recorded in NICU during a painful procedure and the ground truth labels were assigned by a trained nurse with the NIPS scale [22]. In particular, for facial expressions Zamzmi et al. used optical flow, a well-known method of motion estimation, to calculate optical strain magnitudes in the facial tissues of the infants. The prediction was obtained using binary classifiers (SVM and KNN) trained using handcrafted features. An overall accuracy of 95% was achieved by combining all pain indicators for assessment and utilizing LOSOXV as evaluation method. In addition, the different unimodal approaches, where each of the pain indicators is used individually, were compared and the results confirmed that facial expression could be the most specific and common indicator of pain, as it achieved the highest overall accuracy.
This work [54] was then extended to include crying sounds and state of arousal as behavioural measures of pain in [55], achieving an accuracy of 96.6% for the binary classification of the infants' emotional states as no pain or severe pain.

Some recent researches have also combined different modalities, Celona et al. [51] extracted deep features from images of iCOPE dataset with two pre-trained CNNs (VGG-face and MBPCNN) and combined them with LBP and HOG features. The extracted feature vector was used to train SVM to perform binary classification. Testing the trained model on unseen data achieved 82.42%, 81.53%, and 83.78% for VGG-Face, MBPCNN, and VGG-Face+MBPCNN, respectively.

Other authors have therefore applied Convolutional Neural Network models to classify infant pain. In 2018, Zamzmi et al. introduced two distinct techniques based on the use of CNNs. Their first work [56] explored the application of transfer learning using four pre-trained CNN architectures: VGG-F, VGG-S, VGG-M, and VGG-Face. The first three architectures were originally trained on the ImageNet dataset for object classification, while VGG-Face was specifically trained on facial data. They extracted deep features from both high and low layers of these architectures, employing them to train a supervised machine learning classifier. The results demonstrated promising performance, achieving an AUC of 0.841 and an accuracy of 90.34%. It was also observed that the VGG-Face model outperformed others due to its training on a face-specific dataset, leading to a better feature extraction.

More recently, Zamzmi et al. introduced the Neonatal Convolutional Neural Network (N-CNN) [57]. This architecture was designed and trained end-to-end to detect neonatal pain. The N-CNN is structured as a cascaded CNN featuring three convolutional branches. This design enables the combination of image-specific details with general information after applying convolutions. The features generated from these branches are merged and subsequently classified into "pain" or "nonpain" by two fully connected layers. Their N-CNN model demonstrated good performance, achieving an average accuracy of 91% and an AUC of 0.93 on NPAD dataset, along with an average accuracy of 84.5% on the iCOPE dataset.

In 2021, Carlini et al. [58] proposed a mobile application for the Android operating system that uses AI techniques to automatically identify the facial expression of pain in neonates. The framework was based on a CNN architecture, specifically the VGG-Face architecture, fine-tuned on the iCOPE [36] and UNIFESP [53] datasets for binary pain classification. Then the classification model was optimised and embedded in a mobile application. The results showed that the model trained on both datasets lead to better performance (89% cross-validation accuracy) with respect to using only iCOPE (83%) or only UNIFESP (72%) for training. Moreover, an explainable method called Integrated Gradients was employed to the test set of both datasets to identify relevant facial regions for pain assessment. The authors conclude that the nasolabial groove, as well as the open mouth and protrusion of the tongue are likely the most discriminating facial regions to pain assessment.

Conversely, Sun et al. [50] employed Action Units to assess infant pain intensity. The intensities and importance of the AUs were identified through the engagement analysis performed with the Gradient Class Activation Map (Grad-CAM) [59] and were used to train a regression model for infant pain assessment. Experiments were carried out by fine-tuning a pre-trained ResNet model on the YouTube Immunization dataset, and testing on two unseen dataset, i.e. the YouTube Blood Test dataset, and the iCOPEVid dataset. The Grad-CAM layer was added to the ResNet architecture to analyse what the neural networks learn during training. The results showed that the regression model outperformed end-to-end models, and the best model results were achieved by using engagement levels to weight the AUs.

Overall, these studies demonstrated the effectiveness of transfer learning with pre-trained Convolutional Neural Networks for feature extraction in neonatal pain expression recognition, especially with the VGG-Face architecture. Therefore, we selected this approach for our work. Moreover, some studies integrated Explainable AI (XAI) methods to enhance the interpretability of the neural network models and understand which facial regions were considered during the recognition process.

Chapter 3

Deep Learning Theoretical Framework

This Chapter provides a theoretical framework for the thesis, with a particular emphasis on deep learning, Convolutional Neural Networks, the VGG-16 architecture, and methods of explainable artificial intelligence (XAI).

3.1 Overview

The term artificial intelligence (AI) refers to the broad area of computer science that focuses on building intelligent machines capable of performing tasks like learning, problem-solving, and decision-making that normally need human intellect. AI is founded on the concept that machines can be trained to learn from experience in the same way humans do. It includes many different methods, such as Machine Learning (ML) and Deep Learning (DL). Machine Learning is a subset of AI where statistical techniques are employed to enable machines to learn from data without explicit programming. ML algorithms have the ability to automatically learn from experience and make predictions or take decisions about new data. On the other hand, Deep Learning is a subset of ML that processes and analyses data using neural networks. The relationship among AI, ML and DL is illustrated in Figure 3.1 and Deep Learning will be described in more detail in the following Section, since it constitutes the theoretical framework of the approach proposed in this thesis.

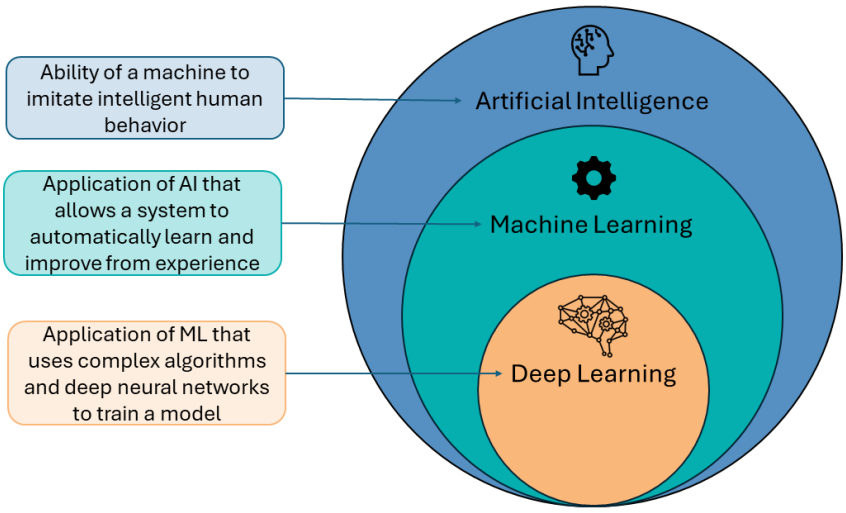


Figure 3.1: Relationship between AI, ML and DL

3.2 Deep Learning

Deep learning comprises a collection of learning techniques aimed at modeling data using complex architectures that integrate various non-linear transformations. The basic elements of deep learning are neural networks, which are assembled to construct deep neural networks. There are two primary learning approaches: supervised and unsupervised techniques. Supervised techniques operate with labeled data, meaning they start with a dataset where each sample has an associated ground truth. Their objective is to establish the relationship between the ground truth and the data samples. Typical examples of supervised learning models include classifiers, which are trained to assign class labels to data samples based on labeled datasets. Conversely, unsupervised learning aims to uncover meaningful patterns within data without relying on pre-existing labels. An example of unsupervised learning is clustering, where data samples are grouped based on inherent similarities rather than predefined categories.

3.2.1 Artificial Neural Networks

Artificial neural networks (ANNs), also known as neural networks (NNs), are mathematical models designed to mimic the behaviour of biological neurons. The architecture of neural networks is characterised by neurons distributed in different layers that exchange information with each other through a complex system of

connections and nodes. As shown in Figure 3.2 every neural network comprises at least three layers:

- An input layer, receiving input data and communicating it to the next layer;
- One or more hidden layers, processing the data to obtain the desired output;
- An output layer, providing the model's results.

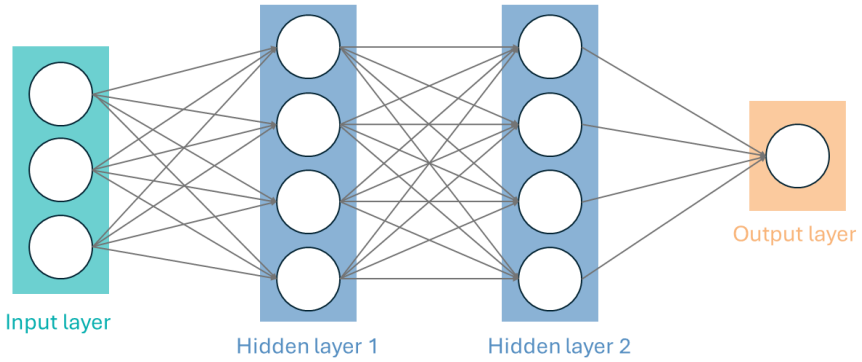


Figure 3.2: Neural Network Architecture

Neurons within each layer are interconnected with those of the subsequent layers by weighted activation functions. The weight is a numerical value that is multiplied by that of the next neuron to determine the importance of any given variable. Each neuron sums the weighted values of all neurons connected to it and possibly adds a bias value. Before passing the data to the next layer, the resulting sum is transformed by applying an activation function. The output of one node then becomes the input of the next node. This neural network is also known as a feedforward network because of this process of passing data from one layer to the next. Following input-to-output propagation, the network evaluates its prediction's accuracy compared to the expected output using a loss function. The network's objective is to minimize this loss by adjusting its weights and biases through training:

$$J(W, b) = \frac{1}{n} \sum_{i=1}^{n} L(\hat{y}_i - y_i) \tag{3.1}$$

where J is the cost function with given weights W and bias b, L is the loss function, $\hat{y}_i$ is the output from the network, and y_i is the expected output.

Activation functions are non-linear functions, which allows the data to be approximated much more precisely. The main activation functions are:

- Sigmoid function: takes a number as input and returns as output a number between 0 and 1. Used mainly in binary classification problems.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

- Softmax: is the generalisation of the Sigmoid function. It is used in multiclass classification algorithms to calculate relative probabilities. It returns the probability of a given input to belong to a given class.

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.3}$$

- TanH (Hyperbolic Tangent): returns an output in the range [-1,1] and unlike the Sigmoid the output is centred on zero.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.4}$$

- ReLu (Rectified Linear Unit): returns an output between 0 and infinity. When the input is positive it returns the input itself as output, if the input is negative it returns 0 as output.

$$ReLU(x) = \max(0, x) \tag{3.5}$$

3.3 Convolutional Neural Networks

Nowadays, Convolutional Neural Networks (CNNs) find extensive application in tasks like image segmentation and classification, object recognition, and face recognition. CNNs are a class of feedforward neural networks specifically designed to process data with a grid-like topology, like images. A digital image is a binary representation of visual information. It is made up of a grid-like arrangement of pixels with values to indicate the color and brightness of each pixel. When an image is perceived, the human brain efficiently processes a huge amount of information. Each neuron operates within its individual receptive field and it is interconnected with others to cover the entire visual field. Similar to how neurons in the biological vision system respond solely to stimuli within their receptive fields, neurons in a CNN exclusively process data within their receptive fields. CNNs are characterised by neurons with three dimensions: height, width and depth and diverge from traditional artificial neural networks in their connectivity pattern. Specifically, the neurons in one layer are only connected to a small region of the previous layer.

This feature enables the processing of very large input images without the need to store an excessive number of parameters.

The CNN architecture consists of three main types of layers: convolutional layers, pooling layers and fully connected layers. Convolutional and pooling layers extract hierarchical visual features from raw pixel data. They start by detecting simple patterns like lines and curves, then progress to more complex ones like faces and objects. On the other hand, the fully connected layers are designated to perform classification or regression.

Convolutional layer The convolutional layer is the core building block of a CNN, responsible for the majority of the network's computational workload. Its parameters primarily focus on the use of learnable kernels, which are small in spatial dimensions but extend through the entire depth of the input. As the kernel slides across the input data during the forward pass, it captures the kernel's response at every spatial position, producing an activation map. This process requires convolving each filter across the input's spatial dimensions, leading to the creation of an activation map. The scalar product between the learnable kernel parameters and a restricted region of the input is computed for each value within the kernel as the input data is scrolled through. Each kernel has a corresponding activation map, which will then be stacked along the depth dimension to compose the complete output volume from the convolutional layer.

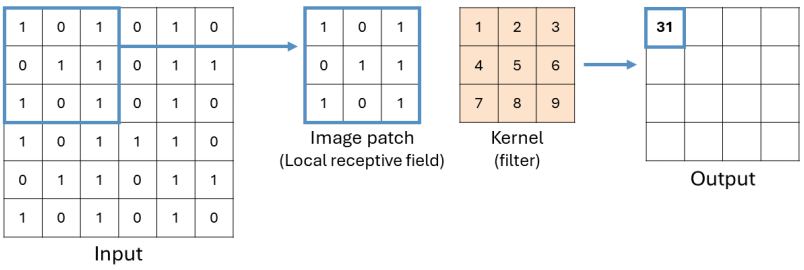


Figure 3.3: Example of convolutional operation

Convolutional layers can also efficiently reduce model complexity by optimizing their output. This optimization relies on three key hyperparameters, which affect the size of the output and need to be set before the network training:

- Depth: the number of filters determines the depth of the output. Reducing this hyperparameter can significantly decrease the total number of neurons in the network, but it can also substantially reduce the model's pattern recognition capabilities.

- Stride: describes the extent to which the kernel traverses the matrix. A stride of 1 results in heavily overlapped receptive fields and large activations, while a larger stride reduces overlap and yields an output with smaller spatial dimensions.

- Zero-padding: process of adding extra values to the border of the input. Setting all elements outside of the input matrix to zero results in an output that is either larger or of equal size.

Each convolutional layer is necessarily followed by an activation function. The purpose of activation functions is to provide the network with non-linear behaviour, enabling the solution of complicated problems using a relatively small number of nodes. A neural network without a non-linear activation function will always behave as a single layer network, regardless of the number of hidden layers present.

Pooling layer The pooling layers enable the reduction of spatial dimensions in the feature map, aiming to maintain crucial information while introducing some degree of translation invariance to the network. Similar to the convolutional layer, the pooling operation applies a filter across the input, but this filter has no weights. Instead, the kernel uses an aggregation function to combine the values within the receptive field, resulting in an output array. These layers usually typically do not have trainable parameters because they only act on small patches of the image summarising local information. However, like convolutional layers, pooling layers also involve hyperparameters such as filter size, stride, and padding. There are two main types of pooling:

- Max pooling: selects the maximum value present in the filter window. This is useful for capturing the most dominant features in a region.

- Average pooling: computes the average value within the filter window. This can be helpful for reducing noise and capturing overall trends.

Figure 3.4 shows an example of operations with max and average pooling.

Fully connected layer The fully connected layers perform the classification task based on the characteristics extracted through the previous layers. Each neuron in this layer is directly linked to the neurons in the two adjacent layers, without any connections to layers within them.

Dropout layer In neural network construction, dropout layers can be added to mitigate overfitting. For each mini-batch in the training set, dropout layer randomly disconnects inputs from the previous layer to the next layer. This

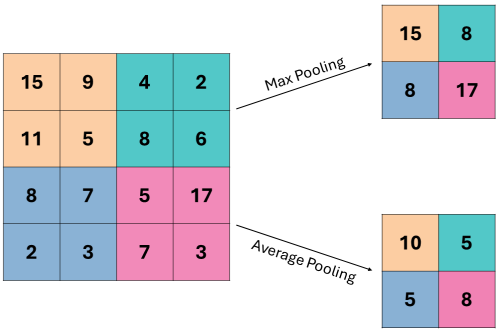


Figure 3.4: Example of pooling operations

operation effectively removes nodes from the network, preventing their influence on predictions and backpropagation. This process creates slightly varied network architectures in each training run, aiding in generalization. The proportion of nodes set to zero is determined by the dropout probability. For instance, a probability of 20% indicates that 20% of the nodes will not be utilised in each run.

3.3.1 VGG-16

In the last decade, several CNN architectures have been developed through structural modifications, regularisation and parameter changes. The architecture plays a crucial role in enhancing the performance of various applications [60]. In particular, popular architectures such as AlexNet, VGG, Inception (GoogLeNet), ResNet and DenseNet have made significant contributions to the field of convolutional neural networks. A detailed description of the VGG-16 network is provided below, as it is the network used in this study.

The VGG-16 model is a convolutional neural network that was proposed for the 2014 ImageNet Challenge by the Visual Geometry Group (VGG) from the University of Oxford [61]. ImageNet is a large image database used in computer vision for object recognition tasks. It contains over 14 million images, each annotated with object labels and bounding boxes. The objects are classified into more than 20,000 categories. In the ImageNet Large Scale Visual Recognition Competition (ILSVRC) a subset of ImageNet with 1000 image categories is used.
Based on research indicating that smaller filter sizes can improve CNN performance, the VGG replaced larger filters with stacks of 3×3 filters in their architecture. They demonstrated empirically that using multiple small size filters concurrently could emulate the effect of a single large size filter. The use of small size filters also offers the advantage of low computational complexity by lowering the number of

parameters [60]. The series of 3 x 3 convolutional layers are stacked on top of each other, for a total of 16 trainable layers, separated by some max pooling layers, which manage the reduction volumes. The max-pooling layers use 2 x 2 filters with a step of 2 in both horizontal and vertical directions. As a result, the output of these layers will be a volume of size $[w_{\text{out}}, h_{\text{out}}, n]$, where $w_{\text{out}} = \frac{w_{\text{in}}}{2}$ and $h_{\text{out}} = \frac{h_{\text{in}}}{2}$, while n remains unchanged. The fact that the spatial dimensions of the representation are reduced makes it possible to reduce the number of parameters and, consequently, also the computational time and the probability of overfitting the data. Overfitting occurs when a model memorises the training data rather than generalising well to unseen data. In addition, each convolution layer is always followed by an activation layer, with the ReLU function. After the series of convolution and max-pooling layers, the VGG-16 network includes three fully connected layers: the initial two have 4096 channels each, while the third contains 1000 channels, each corresponding to a distinct class [61].
The architecture of the VGG-16 network is illustrated in Figure 3.5.

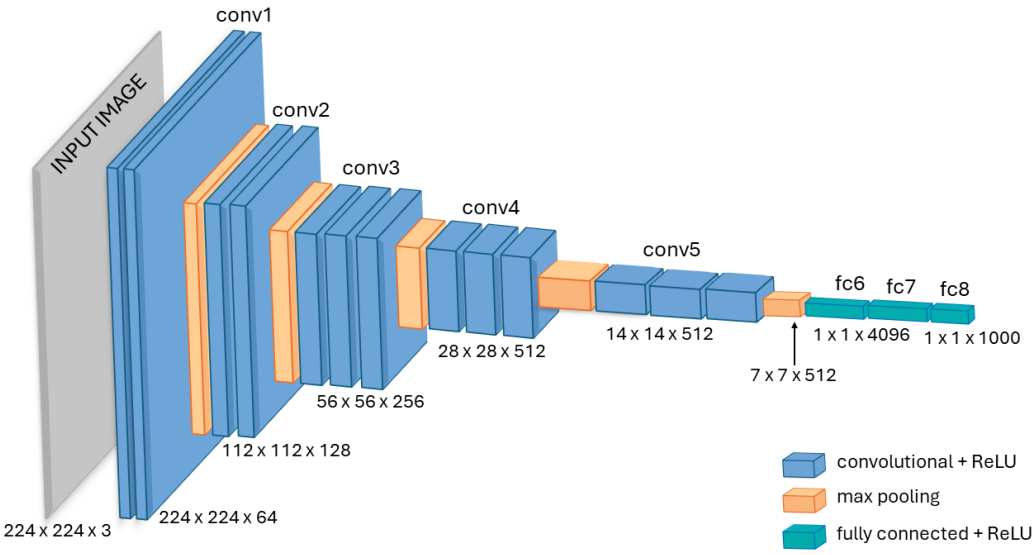


Figure 3.5: VGG-16 model architecture

The VGG-16 model used in this thesis was pre-trained on the first version of the VGG-Face face recognition database [62], which was created by the same research group to classify faces belonging to 2622 individuals. In the VGG-Face architecture, the final fully connected layer consists of 2622 neurons, each corresponding to one of the individuals in the database and producing a value indicating the probability that the input image belongs to that individual.

3.4 Explainability in Deep Learning

Although deep neural networks have revolutionised many domains by achieving exceptional performance across diverse tasks, their inherent complexity and "black box" nature pose significant challenges in terms of interpretability and explainability. Explainable AI methods aim to increase transparency and comprehensibility of decision-making processes in artificial intelligence models, especially in deep learning models. These methods are essential for enabling users to understand how AI model predictions are generated, thereby increasing confidence in the use of such systems, particularly in critical sectors like healthcare [63].
In a recent review [63], the authors underscored the importance of interpretability in healthcare AI applications to increase understanding and trust among clinicians and end-users. They also highlighted the adaptation of general-domain interpretability methods to healthcare problems, which can assist physicians in better understanding data-driven technologies.

Some of the most common Explainable AI methods include:

- LIME (Local Interpretable Model-agnostic Explanations) that generates local explanations for model predictions, identifying the input characteristics that influenced a given prediction.
- SHAP (SHapley Additive exPlanations) that assigns an importance value to each input characteristic, showing how much each variable contributed to the model's prediction.
- CAM (Class Activation Mapping) that is a method for visualising which parts of an image were important for a deep learning model's prediction. CAM is a technique primarily used in computer vision to display the regions of an image that influenced the model's classification.
- RETAIN (Reverse Time Attention Model), which is a method specifically designed for deep learning models in healthcare. It allows for the identification of the parts of temporal data sequences that influenced the model's predictions [63].

3.4.1 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [59] was introduced as a generalization to CAM whose applicability is limited to specific convolutional neural network architectures as it requires feature maps directly preceding the final softmax layer. Grad-CAM is a localization technique that generate visual explanations for any convolutional neural network model without requiring architectural

changes or re-training [59]. It is also class-discriminative, meaning that it is capable of producing a unique visualisation for each class represented in the image.
Figure 3.6 shows examples of Grad-CAM outputs for the "tiger cat" class and "boxer" (dog) class.

(a) Original Image **(b)** Grad-CAM "dog" **(c)** Grad-CAM "cat"

Figure 3.6: Grad-CAM examples. Source: [59]

This technique utilises gradients relative to a class and a layer that requires visualisation. This enables the visualisation of discriminative regions for any convolutional layer, not just the last one. Obviously, the most interesting visualisations are located towards the bottom of the network. Therefore, maps generated from initial convolutional layers are poorly defined, as these layers are the ones containing the smallest filters [59]. The gradient information flowing into the CNN layer enables the identification of which feature maps are activated for a specific class. By calculating a weighted sum of these feature maps, Grad-CAM produces a heat map that highlights the areas in the input image that had the most influence on the model's prediction. Grad-CAM is therefore used post-training on a neural network with fixed weights. An image is fed into the network to compute the Grad-CAM heatmap for a selected class of interest.

Grad-CAM operates through a three-step process, starting from y_c, which represents the score for class c (i.e. the raw output for class c before the softmax):

1 - Compute gradient:

Compute the gradient of y^c with respect to the feature map activations A^k of a convolutional layer

$$\frac{\partial y^c}{\partial A^k} \tag{3.6}$$

2 - Calculate alphas by averaging gradients:

Calculate the neuron importance weights α_k^c by taking the global average pooling of the gradients over the width dimension and the height dimension (indexed by i and j respectively)

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.7}$$

3 - Calculate final Grad-CAM heatmap:

Perform a weighted combination of the feature map activations A^k where the weights are the α_k^c

$$L_{Grad-CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{3.8}$$

The application of the ReLU function to the linear combination of feature maps emphasises positive values while setting negative values to zero. This step aims to highlight features that positively influence the class of interest, in other words the pixels whose intensity should be augmented in order to increase y^c [59].
In addition, it should be noted that the final Grad-CAM heatmap has the same size as the convolutional feature map (14 x 14 in the case of the last convolutional layers of the VGG network), which is smaller in width and height than the original input image. For this reason, the tiny heatmap is up-sampled to match the size of the original image before making the final visualization.

Chapter 4

Materials and Methods

This Chapter describes the construction of the proposed framework for binary pain classification that consists of three main parts:

- We first describe the video data collected in two different clinical environments and annotated with pain score by healthcare professionals.
- We then present a comparative study of face detectors that identify the most suitable tools to extract the face bounding boxes from the video frames.
- Finally, we present the methods used to perform binary pain classification on different datasets and also an explainable AI method that enhance model transparency and interpretability.

4.1 Data Description and Pain Scale Implementation

In this Section, the data used in the study are described. Prior to the start of video recordings for each data collection, informed consent was obtained from participants' parents. Additionally, the study protocols were approved by the Local Ethic Committee.

4.1.1 M-PAIN Dataset

Subjects

The Mauriziano Pain Assessment In Newborns (M-PAIN) dataset is composed of 79 videos of 79 different full term newborns (48 females and 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.

Recording setup

The videos were recorded at the Neonatology department of the Mauriziano Hospital, in Turin, between October 2022 and January 2023. An Intel® RealSense™ depth camera D435i was placed 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, such as the speed of blood outflow and the behavioural state of the child. In addition, in some videos multiple blood samples were taken from the same baby.

Blood sampling procedure

Before being discharged, newborns are subjected to blood tests in order to identify various congenital diseases. Early detection and intervention for these conditions can prevent serious health consequences for the infant, enabling treatment from the earliest days of life and initiation of specialised therapies. Prior to the withdrawal procedure, the infant is placed on the changing table and is wrapped, which is done as a measure of pain containment. It is worth noting that a pacifier is also used to further alleviate the child's discomfort in this process [14]. During the recording, the heart rate and saturation oxygen of the newborn were measured using the Masimo cardiosaturimeter, whose display shows the oxygen saturation value at the top and the heart rate at the bottom. This physiological data were recorded for future investigations, but they were not used in the current study. The parameters were measured by an electrode, which was placed in one foot and held in place with gauze. In the other foot, the nurse took the sample by pricking the heel with a disposable needle and squeezing the foot to encourage blood to flow out (Figure 4.1). Afterwards, the heel was disinfected and a dressing was applied.

Pain score annotation

Each video was assessed by a healthcare professional using the DAN scale, described in the Section 1.2.3. We defined an observation window within each video, which started with the heel squeeze and 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 three contributions: facial expressions, limb movements, and vocal expressions. In this study we consider specifically the facial expressions score, ranging from 0 to 4 and including three parameters: brow bulge, eye squeeze, and nasolabial furrow. Each parameter was individually evaluated and quantified based on the duration of the parameter within the observation window.

Figure 4.1: Example image of M-PAIN dataset during the blood sampling procedure. The image was postprocessed to anonymise infant face.

Subsequently, there was a phase of aggregation of the individual output scores, with the maximum value selected as the final DAN face score. For this study, 76 videos were considered usable after excluding 3 videos where the infant's face was either out of frame or excessively covered by a blanket. Considering only facial expressions, there are 50 videos with a score of 0, 2 videos with a score of 1, 8 videos with a score of 2, 11 videos with a score of 3, and 5 videos with a score of 4. It should be noted that the dataset is imbalanced, with a limited number of videos exhibiting medium or high pain scores.

4.1.2 VideoDol Dataset

Subjects

This dataset is composed of 34 videos featuring 29 different children (14 males and 15 females), aged 3-36 months, admitted to the Pediatric Emergency Department (ED) with acute pain as main or accompanying symptom. Notably, the vast majority of the children are of Caucasian ethnicity, with only one child being of Arabian origin. Children with chronic health conditions, those whose faces and limbs were partially obscured due to dressing, medications, or medical devices, and those assigned high triage priority codes upon admission were excluded from the study.

Recording setup

The videos were recorded at the Regina Margherita Children's Hospital, in Turin, between April 2022 and September 2023. A video capturing the entire body of the children was recorded using an RGB camera with a resolution of 1920 × 1080 pixels and a frame rate of 30 fps. The camera was positioned 100 cm away from the subject, maintaining consistent lighting conditions. Additionally, heart rate and oxygen saturation were evaluated and recorded using the pulse oximeter in the Emergency Department. Figure 4.2 shows an example of the setup.

Figure 4.2: Example image of VideoDol dataset setup. The image was postprocessed to anonymise infant face.

Pain score annotation

Pain assessment was conducted for all children by three different healthcare professionals, using the Italian validated version of the FLACC scale [19] [64], described in Section 1.2.3. However, due to excessive noise in the audio signals recorded in the Emergency Department environment, the cry (C) and consolability (C) categories were excluded from the analysis, as they could not be accurately processed or analysed. In this study we consider specifically the facial expression score, ranging from 0 to 2, which involve mouth opening, brow bulging and eye squeezing parameters. Nevertheless, it is important to note that these parameters were not evaluated individually, unlike the evaluation with the DAN scale. Table A.1 shows the list of videos with the FLACC face score assigned by three different healthcare professionals. It is evident that the majority of the videos have a score of 0. Additionally, in some videos, the ratings of the three operators are inconsistent.

4.2 Comparative Study of Face Detectors

This Section presents a comparative study of face detectors, as the initial step of our approach involved detecting faces to extract bounding boxes from video recording frames.

4.2.1 Face Detection

Face detection is a subset of computer vision technologies that can automatically detect and locate human faces in images or video frames. It is the fundamental initial step in most computer vision applications that involve face analysis. A high-quality facial detector is a necessary starting point for various activities involving the face including facial landmark detection, gender classification, crowd analysis, attendance tracking, face tracking, and facial recognition. Face detection must be robust to various factors, including orientation, lighting conditions, skin tones, hair color, makeup, facial clothing and accessories, age, and so on. The task becomes particularly challenging when applied to young children and newborns, especially in clinical environments. This stems from the fact that many face detection frameworks have been primarily trained on datasets consisting of adult faces [65]. As a result, they struggle to cope with the significant differences in size and proportion that are characteristic of infant faces. Newborns have less well-defined facial features compared to adults, with a wider range of poses and postures, and obstructions like pacifiers, toys, or blankets are very common. Moreover, in the clinical environment, background scenes are complicated, medical equipment and procedures may be present, and varying lighting conditions pose additional challenges [65].
For these reasons, in order to identify the most suitable tool we conducted a comparative study of the face detector on M-PAIN and VideoDol datasets. The following Section provides an overview of selected state-of-the-art face detectors.

4.2.2 State-Of-The-Art Face Detectors

OpenCV

OpenCV's Haar cascade classifier, introduced by Viola and Jones in 2001 [66], represents a traditional approach to face detection. It operates by analyzing Haar features, which are simple rectangular patterns that capture basic differences in brightness and contrast within an image region. By combining these features in various arrangements, the classifier learns to identify patterns characteristic of faces. The classifier is trained using the AdaBoost algorithm, which iteratively selects the most informative features from a large collection of positive (containing faces) and negative (without faces) images. So the classifier scans the image, calculating pixel intensity sums within regions defined by the selected Haar features. If a sufficient

number of features match within a particular region, the classifier identifies that region as containing a face.

Dlib-HOG

Dlib [67] is a C++ library that provides an implementation of the HOG (Histogram of Oriented Gradients) feature descriptor, which can be utilised for tasks such as face detection and object detection. Dlib integrates the traditional HOG feature with a linear SVM classifier model, an image pyramid, and a sliding window detection method. HOG captures edge orientations in image regions, providing a texture and shape representation. The classifier, trained on facial and non-facial images, distinguishes between faces and other objects. The image pyramid allows the detection of faces at different scales, while the sliding window method scans the image at different locations and scales.

SSD

Single-shot multibox detectors (SSD) [68] finds the object in just one iteration through the input image, unlike other models that require multiple passes to generate detection results. The backbone model serves as a feature map extractor, pre-trained for image classification. Stacked convolutional layers form the SSD head, which is added on top of the backbone model. SSD employs a set of default bounding boxes with various aspect ratios and scales for each feature map position. In the face detection process, each of these default boxes produces an output indicating whether a face is detected, and if so, adjusts the box accordingly to fit the detected face.

RetinaFace

RetinaFace [69] is a single-stage face detector that combines the prediction of facial bounding boxes and 2D facial landmark detection. It employs a combination of extra-supervised and self-supervised multi-task learning techniques to perform pixel-wise face localisation in images, even across various scales of face sizes. The model was trained on the WIDER FACE dataset and identifies five key facial landmarks: the right and left eyes, nose, right mouth, and left mouth.

Mediapipe

MediaPipe [70], developed by Google, is a cross-platform framework that provides pre-built components for multimedia pipelines, enabling easy integration for tasks like object detection, face detection and recognition, and pose estimation. Face Detection by MediaPipe is built on BlazeFace [71] a lightweight and efficient

face detection model designed specifically for mobile GPU inference. The feature extraction network is inspired by MobileNetV1/V2, while the anchor scheme is adapted from SSD. MediaPipe enables a rapid facial detection featuring 6 landmarks (right and left eyes, nose, centre mouth, and left and right ears) and the ability to detect multiple faces simultaneously. In addition, it offers two model variants specifically designed to suit different face detection distances. The "long range" model is optimised for detecting faces at medium or long distances from the camera, while the "short range" model is ideal for detecting faces at closer distances.

Yolov5

YOLO (You Only Look Once) is a deep learning system for real-time object detection known for its efficiency and single-stage architecture. In particular, YOLOv5 [72] offers significant improvements in accuracy, speed, scalability, and flexibility compared to its predecessors. YOLOv5 uses pre-trained models to learn from existing data, then combines information from various image scales (low-resolution for coarse details, high-resolution for fine details) to improve detection accuracy for objects of different sizes. Finally, it uses a combined loss function that optimises both object class confidence and precise object location within the image. In our study, we use YOLOv5 object detector [72] with pre-trained weight obtained from [73], which were specifically trained on a subset of the USF-MNPAD-I dataset [45], composed of images of newborns undergoing a painful procedure. In [65] this model has shown better performance for detecting the faces of neonates compared to YOLO model baselines.

4.2.3 Performance Comparison of Face Detectors

To evaluate the performance of the face detectors on the datasets, we extracted frames at a rate of one per second and then we applied the detection methods to these frames for testing purposes. Our evaluation included 76 videos (14148 frames) from the M-PAIN dataset and 30 videos (2491 frames) from the VideoDol dataset. In this comparative analysis two primary metrics were assessed: the percentage of images without detected faces and the corresponding time required for detection. The average execution time per image during the inference phase was calculated using Python on a CPU 2.8GHz 4-Core Intel i7. The results of the comparison between the face detectors are presented in Table 4.1.

On the VideoDol dataset, it is noteworthy that the Mediapipe (long range) method emerged as the top performer, exhibiting the best outcome regarding the proportion of images without detected faces while simultaneously requiring a relatively shorter detection time. RetinaFace showed superior performance in

reducing the number of unrecognised faces, but at the cost of a longer recognition time. Conversely, the other face detection methods exhibited comparatively inferior performance and longer processing times.

On the M-PAIN dataset, many existing methods struggled to detect the newborns' face, reaching 95% of images with no face detected in the case of OpenCV. The best results were obtained with Yolov5, followed by Mediapipe (short range), with a significant difference in time. Yolov5, despite its superior performance, requires much more time for detection compared to Mediapipe.

Overall, these findings confirmed that the majority of commonly used methods demonstrate suboptimal performance, particularly in cases involving newborns with occlusions such as pacifier usage. Selecting the appropriate face detection method (see Sections 4.3.3, 4.3.4) is crucial for achieving optimal results on infant datasets, considering their unique characteristics and challenges. A balance between accuracy and speed is essential when choosing a method for practical applications, for instance a faster method is more suitable for real-time processing.

	VideoDol		**M-PAIN**	
Method	**Empty (%)**	**Time (ms)**	**Empty (%)**	**Time (ms)**
Mediapipe (short range)	25.93	4.6	**18.63**	**3.3**
Mediapipe (long range)	**4.01**	**7.1**	28.53	5.9
OpenCV	22.68	25.4	95.25	20.7
SSD	8.51	22.7	69.42	24.3
RetinaFace	**1.24**	**28.8**	22.26	27.4
Yolov5 with weights	26.69	142.1	**11.12**	**136.7**
Dlib	17.94	238	89.39	106.4

Table 4.1: Results of comparative study of face detectors

4.3 Pain Classification

This Section explains the methodologies used for binary pain classification. Firstly, the general model is explained, which was then applied to two different infant population: newborns undergoing blood tests and young children recorded in the Pediatric Emergency Department. The details of both applications are provided.

4.3.1 Computational Resources and Framework

In order to provide reproducibility and transparency of our work, we present a review of the computational resources and frameworks that we used. Specifically, we used TensorFlow, an open-source library for machine learning and artificial intelligence, along with Keras, a high-level interface that simplifies neural network development. Additionally, we leveraged the computational power of a GPU (Graphics Processing Unit) to accelerate the model training process, enabling us to achieve results in less time and with greater efficiency. In particular, we were equipped NVIDIA GeForce GTX 1060 6GB. The utilised version of Python was 3.10.13 with Scikit-Learn version 1.3.1 and the versions of Tensorflow and Keras were 2.10 with CUDA 11.2.2.

4.3.2 CNN Model

In the following, model architecture and methods used to perform binary pain classification on both infant populations are presented.

Model architecture

For our pain classification model, we decided to use the VGG-Face model, which has been shown to achieve the best performance in literature studies [56] [58]. The VGG-Face network [62], developed by researchers at the Visual Geometry Group (VGG), is based upon the VGG-16 architecture [61] (Section 3.3.1) and it has been pre-trained on 2.6M facial images of over 2.6K people. The network model and respective weights were obtained from [74] that represent the Keras implementations of the models developed by the Visual Geometry Group.

To adapt the model to our requirements, we hence customised the top layers of the architecture. We substituted the last flattening and fully connected layers with the average pooling layer and three dense layers, the first two consisting of 2048 and 1024 neurons, respectively, both employing ReLU activation functions. In addition, dropout layers with a 50% rate were inserted after each dense layer to mitigate overfitting. The final dense layer consists of two units with the Softmax activation function, generating binary outputs for *pain* or *nonpain* classes. While a simple Sigmoid can also be effective for binary classification, the Softmax performs better as an output layer. The Sigmoid produces only a single probability for the positive class. On the other hand, Softmax's output probabilities are interrelated and always add up to 1; consequently, a higher output value for one class leads to a lower output for the other class.

Transfer learning

Given the limited size of the datasets used in this study, a transfer learning approach was adopted. The main task associated to the VGG-Face network is face recognition. Consequently, the initial layers of the VGG-Face model are able to capture low-level features and general facial representations. While this provides a good starting point, in order to adapt the acquired features to our specific problem, it was necessary to further fine-tune the network on the infants' data. During the training process, the first 15 layers were kept frozen, while the parameters of the last group of convolutional layers were updated to capture more specialised facial features of infants. The resulting architecture had 10 230 274 trainable parameters and it is illustrated in Figure 4.3.

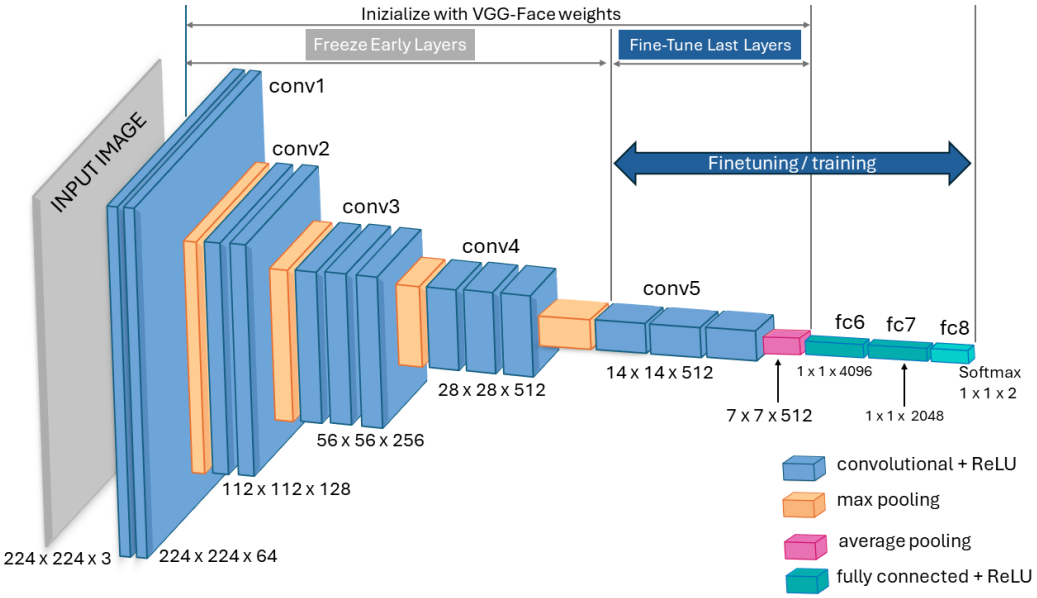


Figure 4.3: CNN model architecture

Data preprocessing

To maintain consistency with the training procedures of the VGG-Face pre-trained model, the same preprocessing method used in the VGG-Face dataset [62] was employed. Thus, the images were converted from RGB to BGR, and each color channel was zero-centered with respect to the mean values of the VGG-Face dataset. All the images were resized to 224 $\times$ 224 pixels, to match the expected dimensions of the pre-trained model. An overview of the complete training CNN process is depicted in Figure 4.4.

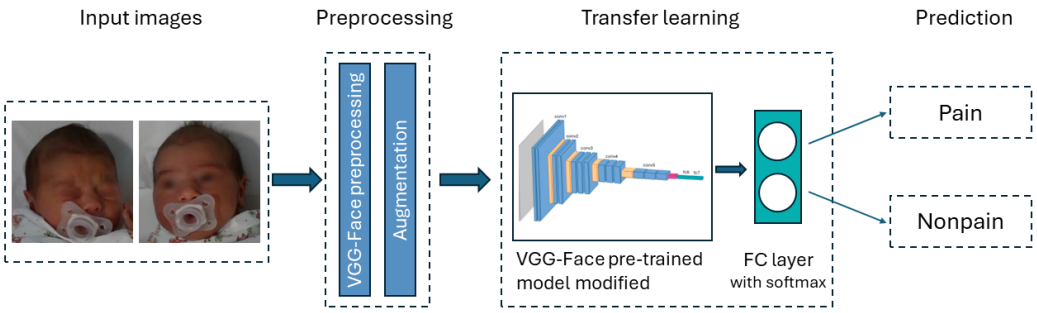


Figure 4.4: Overview of the complete training CNN process

Class imbalance

Class imbalance is a prevalent characteristic across all datasets utilised for training our CNN models, with a notably higher number of samples belonging to the *nonpain* class compared to the *pain* class. To address this issue, we included class weights to ensure that the model gives more attention to instances from the underrepresented class. Specifically, the Scikit-Learn library was used to compute class weights which were then incorporated into the fit() function as a parameter in order to weight the loss function during training. Although our task was binary classification, we used categorical cross-entropy as the loss function. This choice is motivated by the relationship between categorical cross-entropy and the softmax function, which will be discussed in more detail later.

Categorical Cross-Entropy Loss Function:

$$\mathcal{L}_{CCE} = -\mathbb{E}\left[\sum_{i=1}^{C} y_i \cdot log(\hat{y}_i)\right] \tag{4.1}$$

Weighted Categorical Cross-Entropy Loss Function:

$$\mathcal{L}_{wCCE} = -\mathbb{E}\left[\sum_{i=1}^{C} w_i \cdot y_i \cdot log(\hat{y}_i)\right] \tag{4.2}$$

where:

– C is the number of classes

– y_i is the true label

– $\hat{y}_i$ is the model prediction

– w_i is the class weight

Performance parameters

In order to evaluate the performance of the model, the important performance parameters, i.e., accuracy, precision, recall, specificity, and F1-score were calculated using Equations (4.3). The values needed to calculate the performance parameters mentioned above, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) were derived from confusion matrices (Table 4.2).

$$\text{Precision} = \frac{nTP}{nTP + nFP} \tag{4.3a}$$

$$\text{Recall} = \frac{nTP}{nTP + nFN} \tag{4.3b}$$

$$\text{Specificity} = \frac{nTN}{nTN + nFP} \tag{4.3c}$$

$$\text{Accuracy} = \frac{nTP + nTN}{nTP + nTN + nFP + nFN} \tag{4.3d}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3e}$$

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 4.2: Confusion matrix illustrating the model predictions against true class labels

In our binary classification task, the *nonpain* and *pain* images were considered as negative and positive cases, respectively. Thus, nTN and nTP denote the correctly predicted *nonpain* and *pain* images, while nFN and nFP represent the falsely predicted *nonpain* and *pain* images, respectively. Because of class imbalance, the

F1-score metric was chosen as an evaluation metric in addition to the accuracy. In this way, it was possible to measure the model's ability to correctly identify images depicting pain while minimising false positives.

4.3.3 CNN for Pain Classification in Newborns

This Section describes the experiment conducted on the M-PAIN dataset (Section 4.1.1) and iCOPE dataset. The same model architecture and methods described in Section 4.3.2 were used, while the remaining implementation details are presented in the following paragraphs.

Frame extraction and labelling on M-PAIN dataset

In order to build our neural network training dataset, we implemented a systematic frame extraction process from videos (cut on the observation window as described in Section 4.1.1), aimed at preventing a significant class imbalance. For videos with a DAN facial score of 0 or 1, one frame was selected every two seconds of footage. Conversely, for videos with score between 2 and 4, we opted for a higher extraction rate, specifically one frame per second. Since the length of the observation window was variable, for the longer videos only the first part of the observation window was 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 expression. Following this considerations, we manually labelled all the extracted frames under the supervision of a healthcare professional. Frames from videos with a pain score of 0 or 1 were categorised as belonging to class *nonpain*; on the other hand, frames from videos with scores ranging from 2 to 4 were classified as belonging to class *pain* or class *nonpain* according to the facial expression present in each frame, as evaluated by the healthcare professional.

Face detection

The results of the face detector comparison, presented in Section 4.2.3, showed that on the M-PAIN dataset Yolov5 with pre-trained weights [73] performed best, being able to identify the faces of newborns in most cases. For this reason we employed this methods to detect and crop the subject's face in order to obtain the facial images to train the network.

Model details

At the end of the preparation process of the M-PAIN dataset we obtained a total of 5309 frames, with 3934 labelled as *nonpain* and 1375 labelled as *pain*. While the iCOPE dataset is composed of 204 images, with 144 labelled as *nonpain* and the remaining labelled as *pain*. In both cases, we divided each dataset in three different sets: training, validation and test set.

Experiments performed Three experiments were performed. In the first experiment, we built a pain classification model using only the M-PAIN dataset. In the second experiment, we used the same approach on the iCOPE dataset, to enable comparisons with similar research works. Finally, we performed model training on both the M-PAIN and iCOPE dataset, to check if the performance on the M-PAIN data can be further improved by the combination of different training data.

Data augmentation To expand the data and enhance its variability data augmentation techniques were applied. However, data augmentation techniques should be applied in moderation to avoid introducing unrealistic or implausible variations that deviate too far from the real-world distribution. We limited data augmentation to include only random horizontal flips, random rotations within a 30-degree range, and brightness adjustments in the range (1.0 - 1.5). These techniques were applied to the images in both training and validation sets, shuffling the data, along with VGG-Face preprocessing; on the other hand, for the test set images, only the VGG-Face preprocessing step was applied.

Regularisation and hyperparameters With the aim of mitigating overfitting, besides adding dropout layers, other regularisation techniques are employed. We included early stopping to terminate the training when the validation loss fails to improve for 15 consecutive epochs and we employed the categorical cross-entropy loss function along with L1 regularization penalty. In order to mitigate the problem of class imbalance we employed class weights to adjust the loss function.
In all the three experiments performed, the Adam optimizer was used. This optimizer was chosen because, unlike SGD, it is unaffected by the scaling change induced on the loss by class weights [75]. The batch size was set to 32, which was the largest feasible value given our available computing resources, and the maximum number of epochs was set to 100. For what concerns the initial learning rate, it was set to 5e-6 in the first and third experiments, which involve the M-PAIN dataset. While in the second experiment, involving iCOPE, a larger value was chosen (1e-5) to speed up convergence.

Model evaluation To evaluate the trained model and estimate its generalization performance, k-fold cross-validation was employed to avoid the limitations of a single train-validation-test split, which may not adequately represent the overall population, particularly for small datasets. In particular we use StratifiedGroupK-Fold, provided by the scikit-learn library in Python [76]. This technique creates folds that preserve the proportion of samples for each class and ensure that no groups overlap 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 was partitioned in three parts: training, validation and test sets. A model was trained on the training set, while the test set performance was computed using the model weights obtained on the epoch with the highest validation accuracy. The final cross-validation performance was obtained by averaging the performance of the k trained models. In the first and third experiments, with M-PAIN data, k was set to 5. In this way, each fold had a test set comprising approximately 20% of the total dataset, while the validation set was then derived from the training set and comprises approximately 20% of it. The second experiment, instead, involves only iCOPE data, so k was set to 10, given the smaller size of the dataset. Figure 4.5 shows the StratifiedGroupKFold schema for the first experiment using only the M-PAIN dataset, in our case the groups denotes individual newborns, with each newborn being associated with a distinct set of frames.

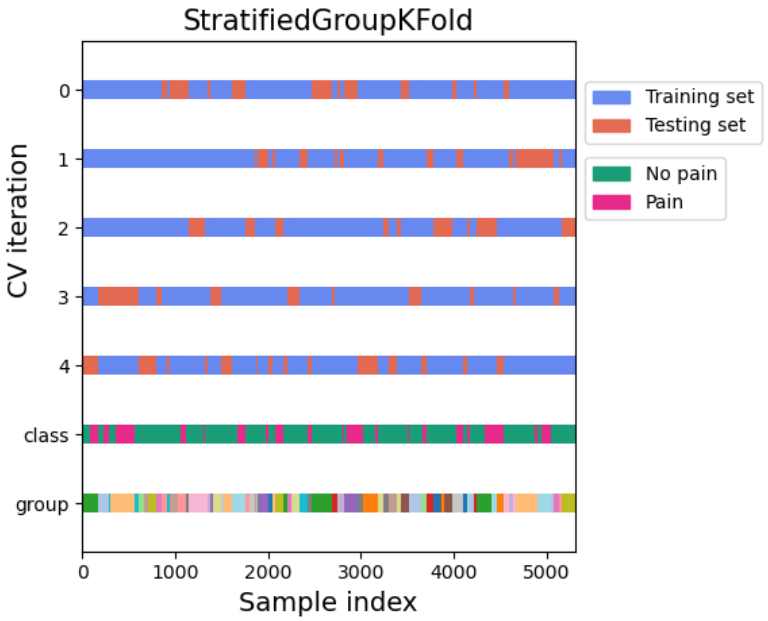


Figure 4.5: StratifiedGroupKFold on M-PAIN dataset (training set includes both training and validation data).

4.3.4 CNN for Pain Classification in Pediatric ED

This Section describes the experiment conducted on the VideoDol dataset (Section 4.1.2) that involve young children, aged 3-36 months, admitted to the Pediatric ED with acute pain. The same model architecture and methods described in Section 4.3.2 were used, while the remaining implementation details are presented in the following paragraphs.

Frame extraction and labelling

Following a similar methodology to that described in the previous Section (4.3.3), we employed a predefined strategy for extracting frames to avoid an excessive class imbalance in the dataset. From the VideoDol data described in Section 4.1.2 we excluded 3 videos to prevent redundancy, as there were already existing videos with low pain assessment scores featuring those specific children. Furthermore, considering the varying lengths of the videos, to maintain balance and prevent the dataset from being skewed by longer videos with low pain scores, we chose to analyze only the initial approximately one-minute segments for such cases. We selected frames at a rate of one frame per second for videos with a FLACC face score equal to 0 or 1, while for videos with a score of 2 we chose a rate of two frames per second. Under the supervision of a healthcare professional, we have labelled the frame in this way: frames belonging to videos with a pain score of 0 or 1 were categorised as belonging to class *nonpain*; frames belonging to videos with score equal to 2 were classified as belonging to class *pain*. In consideration of some differences in the scores estimated by different healthcare professionals, we decided to eliminate the frame associated to one video with inconsistent facial pain scores, equal to 1 or 2. At the end we considered 30 videos of 28 different children. It is crucial to highlight that only 5 videos, belonging to 3 different children, had a score of 2. This significant class imbalance, coupled with the scarcity of *pain* images, posed a challenge in training the network effectively.

Face detection

Based on the findings of our comparative analysis of face detectors (Section 4.2.3), we chose to use Mediapipe for the detection and cropping of the infant's face within each frame. This choice stems from Mediapipe's ability to offer extensive coverage in face detection, thus ensuring both sufficiently high precision and fast execution times. Afterward, we conducted a manual review to exclude some frames where the face was excessively covered, whether by objects, the operator's hands, or the child's hands.

Model details

At the end of the frame extraction and face detection phases we obtained a total of 2026 frames, with 1414 labeled as *nonpain* and 612 frames labeled as *pain*. Considering the small number of samples, especially of frames classified as *pain*, belonging to only 3 different children, it was not feasible to divide them into training, validation and test sets. Consequently, we opted for a single division into training (75%) and test (25%) sets, maintaining in each set the same proportion of *pain* and *nonpain* images of the entire dataset. To mitigate the issue of class imbalance, we employed the class weight (Section 4.3.2).

Data augmentation Data augmentation techniques were applied to the training data to increase its volume and improve variability, along with VGG-Face preprocessing. We limited image data augmentation to include only random horizontal flips and random rotations within a 30-degree range.

Regularisation and hyperparameters We employed the categorical cross-entropy loss function along with L1 regularization penalty. We used the Adam optimizer with a learning rate equal to 5e-6. This optimizer was chosen because, unlike SGD, it is unaffected by the scaling change induced on the loss by class weights [75], which was employed to address the issue of class imbalance. The batch size was set to 32, which was the largest feasible value given our available computing resources.

Model evaluation To address the risk of overfitting, we restricted the training duration to a limited number of epochs, i.e. 20, ensuring that the model did not excessively specialise on the training data. This approach aimed to promote generalization by preventing the model from memorising the training samples. Subsequently, to assess the model's performance, we employed the model parameters obtained from the final epoch and evaluated its predictive capabilities on the independent test set.

4.4 Explainability

To provide model explainability and transparency, we used an explainable AI method, i.e. Gradient-weighted Class Activation Mapping (Grad-CAM) [59]. Grad-CAM is a visualisation technique, which is used to create a class-specific heatmap that allow to understand which parts of an image the network focuses on when making a prediction. Figure 4.6 shows a schematic diagram of the Grad-CAM method for generating the heatmap from an input image. In our implementation,

the inputs specified to the Grad-CAM algorithm were the trained classification model, the final convolutional layer, together with the test images and the corresponding model predictions. Compared to initial convolutional layers, which have smaller receptive fields and exclusively focus on local features, later layers better capture high-level semantic information while keeping spatial information [59]. In fact, the final convolutional layer is the nearest layer to the classification target that preserves spatial information essential for capturing visual patterns.

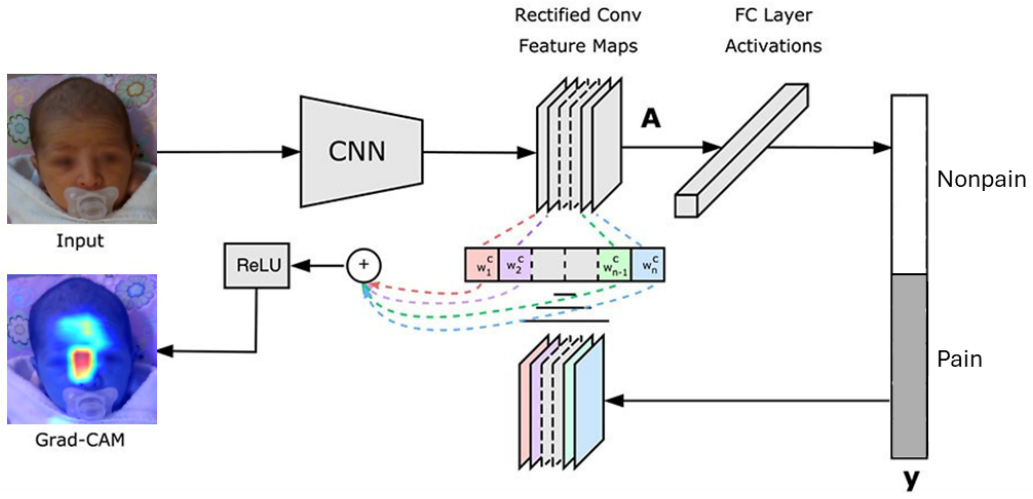


Figure 4.6: Architecture to describe the Grad-CAM technique for generating the heatmap from an input image. Adapted from [59]

Chapter 5

Results and Discussion

In this Chapter, the results of the infant pain assessment approach described in Section 4.3 are presented and discussed. The following Section presents the results of the experiments conducted on the M-PAIN and iCOPE datasets, followed by the results obtained on the VideoDol dataset. In addition, examples of the resulting Grad-CAM heatmaps on the test set images are provided, where the heatmaps have been simplified to ensure a clearer presentation of the underlying image.

5.1 CNN for Pain Classification in Newborns

For newborn pain assessment three different models were trained, following the approach described in Sections 4.3.2, 4.3.3. The results in terms of mean and standard deviation of the accuracy and F1-score over the cross-validation steps are shown in Table 5.1.

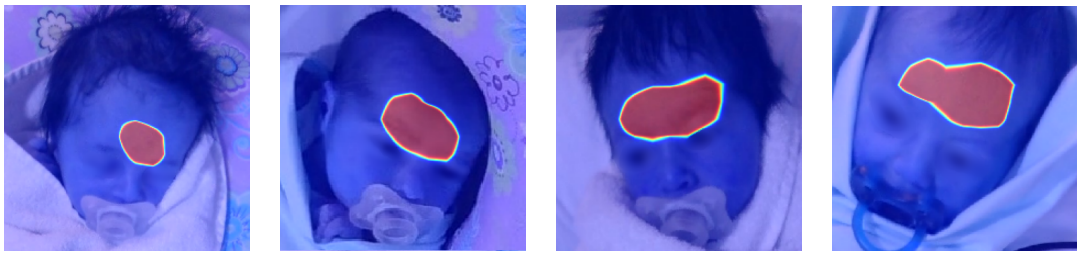
The first model, trained and tested on the M-PAIN dataset using 5-fold cross-validation, achieved an average accuracy of 87.4% ± 3.4% in classifying *pain* vs *nonpain* images. Due to the significant class imbalance within the dataset, with

Experiments performed		Accuracy	F1-score
1 -	Training and evaluation on M-PAIN	0.874 ± 0.034	0.754 ± 0.054
2 -	Training and evaluation on iCOPE	0.838 ± 0.107	0.668 ± 0.293
3 -	Training on M-PAIN+iCOPE, evaluation on M-PAIN	0.888 ± 0.050	0.796 ± 0.081

Table 5.1: Cross-validation results for *pain* vs *nonpain* classification, in terms of accuracy and F1-score (mean ± standard deviation).

approximately 74% of examples representing *nonpain* instances and 26% depicting *pain* instances, the F1-score metric was also considered. This metric evaluates the model's performance in accurately detecting images depicting *pain* (considered positive examples), while also mitigating the occurrence of false positives. The effectiveness of the model in detecting pain images is demonstrated by the average F1-score of 75.4% ± 5.4%.
Upon analysing the confusion matrices of each cross-validation fold, it was observed that the model exhibited different behaviours. In some cases, the model accurately classified almost all images labelled as *nonpain*, but struggled more to classify *pain* images. Conversely, in other cases, the model appeared to perform better at classifying *pain* images than *nonpain* images. The differences among the various folds may be due to variations in the composition of the training, validation, or test data, as well as to the fact that the model itself was saved based on its performance on the validation set. Each fold may contain a different distribution of examples, with variations in the quality or representativeness of the data itself. For instance, one fold test set may have more challenging cases than others.
Another important result emerged from the analysis performed through Grad-CAM, in which images from the M-PAIN test set of each cross-validation step were considered. Some examples of the resulting heatmaps are shown in Figure 5.1, where the areas in red are the ones that have most influenced the prediction of the model. The forehead emerged as the most highlighted facial feature in images classified as *nonpain* (Figures 5.1a, 5.1b, 5.1c, 5.1d). Conversely, for images classified as *pain* (Figures 5.1e, 5.1f, 5.1g, 5.1h), Grad-CAM highlighted the upper contour of the nose, the eye squeeze, and the nasolabial groove. Notably, the gradients did not emphasise the pacifier or secondary artifacts such as the blanket, focusing primarily on the newborn's face.
Despite the inability to analyse the mouth due to the presence of the pacifier, the network was able to detect pain by focusing on the other facial regions most closely associated with the experience of pain, in accordance with pain scales commonly used in clinical practice. On the other hand, analysing the frames where infants were highly agitated and lose their pacifiers, it becomes evident that in some cases the network redirected its focus also to the mouth, as shown in the Figure 5.2.

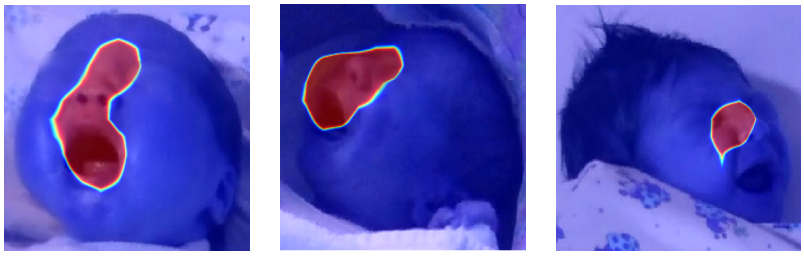
The second model was trained on the iCOPE dataset using 10-fold cross-validation and it achieved an average accuracy of 83.8% and average F1-score of 66.8%. These results exhibit a high standard deviation, indicating significant performance variation across different folds. Such variability is typical when dealing with small dataset sizes, as is the case with iCOPE. Moreover, iCOPE dataset is known to be highly challenging due to the presence of *nonpain* images induced by stressful stimuli, such as friction or moving to a new crib, which in some cases can provoke the infant's crying. Our results on binary pain classification with iCOPE



(a) No pain **(b)** No pain **(c)** No pain **(d)** No pain

(e) Pain **(f)** Pain **(g)** Pain **(h)** Pain

Figure 5.1: Examples of resulting Grad-CAM heatmaps on M-PAIN test set images. Labels are reported below each image. The images were postprocessed to anonymise infant faces.

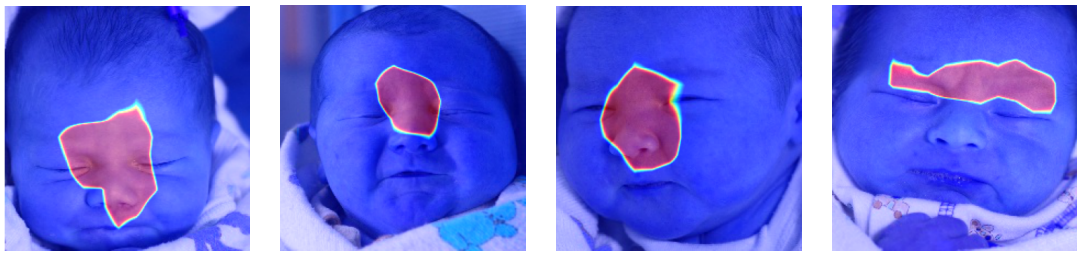


(a) Pain **(b)** Pain **(c)** Pain

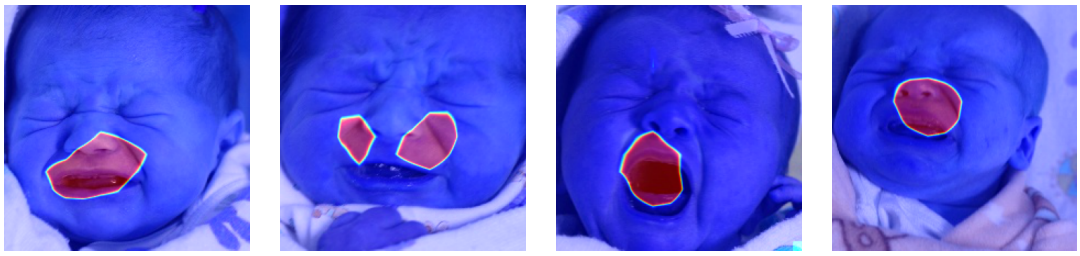
Figure 5.2: Examples of resulting Grad-CAM heatmaps on M-PAIN test set images of newborns without pacifier. Labels are reported below each image. The images were postprocessed to anonymise infant faces.

are comparable to state-of-the-art results in prior studies [37] [51] [57] [58]. However, it is important to recognise that the comparison of these related research works and our model is also based on different evaluation methods. Leave-One-Subject-Out (LOSO) method used in other studies [38, 51] may overestimate accuracy on iCOPE dataset, 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. Moreover, it is noteworthy to highlight the differences between the iCOPE dataset and the M-PAIN dataset. Firstly, the presence of the pacifier in the M-PAIN dataset, and secondly, the different data acquisition and labelling protocol. Figure 5.3 shows the resulting Grad-CAM heatmaps on iCOPE test set. For images classified as *pain* (Figures 5.3e, 5.3f, 5.3g, 5.3h) Grad-CAM highlighted the nasolabial groove and the open mouth. Conversely, the forehead and the nose area emerged as the most highlighted facial features in images classified as *nonpain* (Figures 5.3a, 5.3b, 5.3c, 5.3d).



(a) No pain **(b)** No pain **(c)** No pain **(d)** No pain

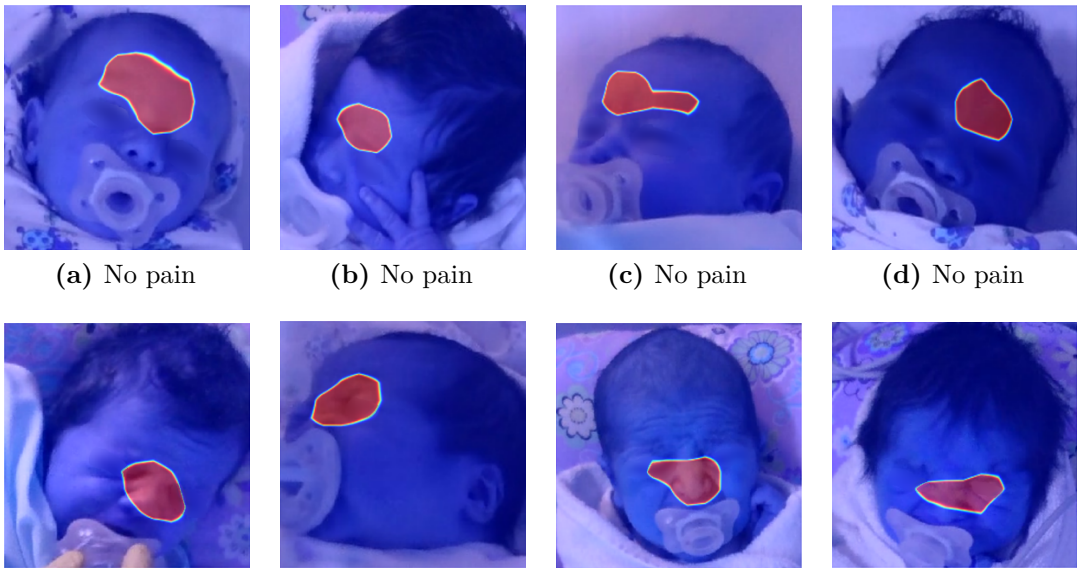


(e) Pain **(f)** Pain **(g)** Pain **(h)** Pain

Figure 5.3: Examples of resulting Grad-CAM heatmaps on iCOPE test set. Labels are reported below each image.

Finally, the third model was trained on both M-PAIN and iCOPE datasets, and tested on M-PAIN. This cross-database training led to an improvement in both accuracy and F1-score compared to the first model, which was trained only on M-PAIN dataset. This suggests that incorporating a wider range of training data can enhance performance, particularly in challenging scenarios. The Grad-CAM heatmaps for M-PAIN test set images relative to this third experiment are shown in Figure 5.4. Again, as in the first experiment, we can see that the facial regions most closely associated with the experience of pain are highlighted. The forehead

emerged as the most highlighted facial feature in images classified as *nonpain* (Figures 5.4a, 5.4b, 5.4c, 5.4d). On the other hand, in images classified as *pain* (Figures 5.4e, 5.4f, 5.4g, 5.4h), Grad-CAM highlighted the upper contour of the nose, the eye squeeze, and the nasolabial groove.



(a) No pain **(b)** No pain **(c)** No pain **(d)** No pain

(e) Pain **(f)** Pain **(g)** Pain **(h)** Pain

Figure 5.4: Examples of resulting Grad-CAM heatmaps on M-PAIN test set images with training cross-database. Labels are reported below each image. The images were postprocessed to anonymise infant faces.

5.2 CNN for Pain Classification in Pediatric ED

The model, trained on frames extracted from VideoDol dataset with a single train-test split as described in Sections 4.3.2, 4.3.4, achieved an accuracy of 88.8% and a F1-score of 78.8%. It should be noted that the model demonstrated higher accuracy in predicting images associated with *nonpain* compared to those associated with *pain*. This discrepancy could be attributed to the limited number of training examples of *pain* images, belonging to only two different infants. Due to the relatively small dataset size for pain-related images, the model may not have had enough exposure to generalise patterns associated with pain, resulting in a performance bias towards *nonpain* images. The normalized confusion matrix is represented in Figure 5.5. Given the absence of false positives, precision reaches its maximum value of 1.00, indicating the model's ability to accurately identify all

nonpain cases. However, with a recall of 0.65, the model demonstrates a moderate performance in capturing *pain* images relative to the total number of actual *pain* cases.

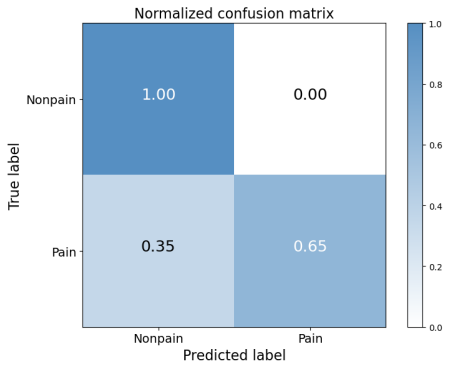
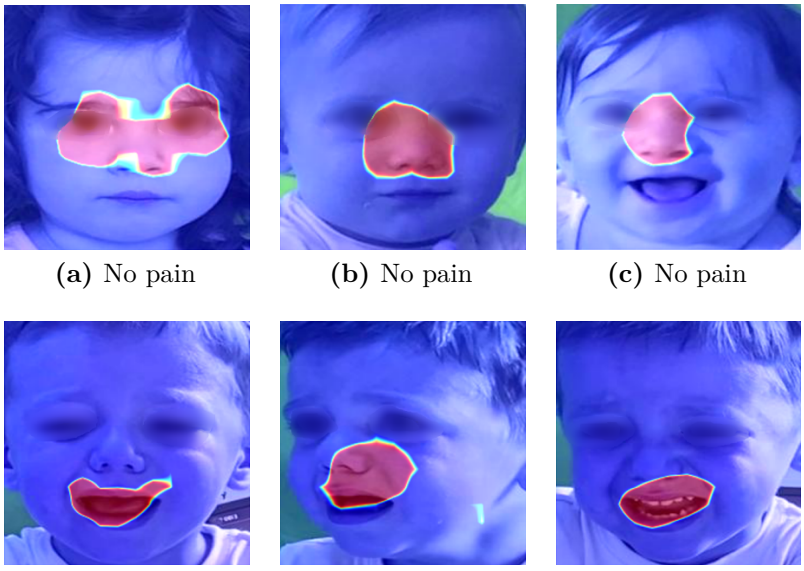


Figure 5.5: Normalized confusion matrix of VideoDol test set.

The analysis performed through Grad-CAM (Figure 5.6) shows that the most highlighted facial feature in images correctly classified as *nonpain* are the nose area and the eyes. Conversely, for images correctly classified as *pain*, Grad-CAM highlights the open mouth and the nasolabial groove.



(a) No pain **(b)** No pain **(c)** No pain

(d) Pain **(e)** Pain **(f)** Pain

Figure 5.6: Examples of resulting Grad-CAM heatmaps on VideoDol test set images correctly classified. Labels are reported below each image. The images were postprocessed to anonymise infant faces.

For some misclassified images, when the correct label, i.e. *pain*, is passed to Grad-CAM, the mouth remains the most prominently highlighted facial feature (Figures 5.7b, 5.7d). However, generating the heatmap based on the predicted, and thus incorrect, label Grad-CAM highlights the nose or other irrelevant areas. (Figures 5.7a, 5.7c).

(a) No pain **(b)** Pain

(c) No pain **(d)** Pain

Figure 5.7: Examples of resulting Grad-CAM heatmaps on misclassified test set images of VideoDol dataset. Labels passed to Grad-CAM are reported below each image. The images were postprocessed to anonymise infant faces.

5.3 Results Discussion and Limitations

Overall, our promising findings pave the way for the development of an automated system that integrates, standardises, and improves human pain evaluation. The use of a visual explanation technique further validates our framework, while revealing interpretability aspects which can provide the basis for an effective integration of automatic pain assessment in the clinical practice. Furthermore, it promotes better understanding and trust among healthcare professionals. It is important to note that the M-PAIN dataset aims to overcome some limitations of other literature datasets, as a small number of subjects, and inadequate annotations or protocol documentation [24] [43]. This dataset reflects the typical characteristics of the real-world environment of the Neonatology Department. It includes the adoption of pain management strategies, such as the pacifier, diverse lighting conditions,

and dynamic newborn positions, including variations in facial orientation. The availability of numerous and diverse data has facilitated the development of a reliable and effective framework.

On the other hand, the Videodol dataset provides a realistic context, but its limited size hinders the ability to train a neural network in a robust manner, which in turn prevents effective generalisation to unseen data during training. Moreover, it is worth noting the substantial disparity between the M-PAIN and VideoDol datasets. These datasets originate from distinct contexts characterised by different setups. The M-PAIN dataset was captured in the Neonatology Department during procedural blood sampling, where pain management strategies such as pacifier use are present. Conversely, the VideoDol dataset encompasses observations of infants up to three years old admitted to the Pediatric Emergency Department, where pacifier use is less common. Therefore, the datasets also differ in the types of pain, with one focusing on procedural acute pain and the other on acute pain. In fact, given their significant dissimilarity, it was not feasible to merge these two datasets.

In addition to the limited data availability within the VideoDol dataset, this study acknowledges several other limitations. One notable limitation is the demographic homogeneity observed among the subjects, who are predominantly individuals of Caucasian ethnicity. This homogeneity may introduce bias and hinder the generalisation of the findings to more ethnically diverse populations. Moreover, relying exclusively on static images for analysis presents a significant constraint. By limiting the analysis to static frames, the complex temporal dynamics inherent in facial expressions are overlooked. This prevents a comprehensive understanding of the evolving nature of pain expression, thereby limiting the depth of insights that could be gained from the data. Furthermore, binary classification, although efficient for simplifying analytical processes, does not capture the nuanced spectrum of pain intensity. Pain, as a multifaceted and subjective phenomenon, encompasses varying degrees of severity that are inadequately captured by binary classification. As a result, the study may overlook subtler nuances in pain expression that could contribute to a more comprehensive understanding of the phenomenon. Addressing these limitations by incorporating more diverse subject populations, considering the temporal dynamics of facial expressions, and exploring methods for capturing the intensity of pain expression could enhance the robustness and applicability of future studies in this area.

Chapter 6

Conclusion and Future Works

This study presents a deep learning framework for pain assessment in infants based on facial expression analysis. It was developed using frames extracted from two different video datasets recorded in real-world conditions. The first (M-PAIN dataset) included newborns undergoing heel stick procedures for blood sampling, recorded in the neonatal unit of the AO Ordine Mauriziano Hospital, while the second (VideoDol dataset) included healthy children, aged 3-36 months, admitted with acute pain to the Pediatric Emergency Department of the Regina Margherita Hospital.

The model trained with a transfer learning technique on the M-PAIN dataset reached an average accuracy of 87.4% and average F1-score of 75.4% using a stratified 5-fold cross-validation. In this setup, pain management strategies as the use of a pacifier were adopted, making the analysis of newborn facial expressions more challenging. In addition, performance on M-PAIN data improved, reaching an average accuracy of 88.8% and an F1-score of 79.6%, when the training was carried out adding images from another existing neonatal dataset, i.e. iCOPE. The same transfer learning approach was used with the image data recorded in the Pediatric Emergency Department. The model trained on the VideoDol dataset achieved an accuracy of 88.8% and a F1-score of 78.8%. Moreover, the application of an explainable AI method based on Grad-CAM showed that the networks' decisions are based on the facial regions most closely associated to the experience of pain in infants, i.e., brow bulge, eye squeeze, and nasolabial furrow.

In conclusion, our promising findings suggest the potential for developing an automated system that integrates, standardises, and improves human pain evaluation.

The results shows that automated pain detection from facial expressions is feasible even in real-world settings. The use of a visual explanation technique enables the interpretability of neural networks outcomes, thereby increasing transparency and trust among healthcare professionals. Furthermore, this research work paves the way for further developments by providing a starting point for subsequent investigations into automatic infant pain assessment.

Several possible directions for further research can be identified:

- Model Enhancement: this may involve investigating alternative neural network architectures, exploring the robustness of our framework against artefacts due to occlusions or movements, and expanding training datasets with additional data.
- Model Extension: the framework could be extended into a multiclass classification task to infer pain intensity levels or it could incorporate additional data, such as vital signs or body motion analysis, where feasible. Additionally, future work could include the investigation of the temporal dimension of pain exploiting videos to gain additional insights.
- Real-Time Pain Detection: future applications of the model could include real-time pain assessment, enabling improved patient monitoring and informed decision-making for better pain management. Such a system could also be useful in more critical settings, like Neonatal Intensive Care Units (NICUs), where continuous monitoring and early pain detection are crucial for improving outcomes and reducing suffering in vulnerable neonates.

Appendix A

Evaluations

A.1

		FLACC Face Score		
Video ID	**Infant ID**	**Operator 1**	**Operator 2**	**Operator 3**
0	173047	0	0	0
1	173047	0	0	0
2	181140	0	0	0
3	4163	0	0	0
4	4228	0	0	0
5	4295	0	0	0
6	5158	0	0	0
7	5242	0	0	0
8	5244	0	0	0
9	5244	0	0	0
10	5246	0	0	0
11	8159	0	0	0
12	8201	0	0	0
13	8517	0	0	0
14	8628	0	0	0
15	8631	0	0	0
16	8632	0	0	0
17	8669	0	0	0
18	8676	0	0	0
19	8686	0	0	0
20	8688	0	0	0
21	5191	1	0	0
22	8531	1	0	0
23	8276	0	1	0
24	8523	0	1	0
25	8565	0	1	1
26	4163	1	1	1
27	4226	1	1	1
28	8203	1	2	1
29	3759	2	2	2
30	4155	2	2	2
31	4155	2	2	2
32	4203	2	2	2
33	4203	2	2	2

Table A.1: FLACC face score for VideoDol dataset

Bibliography

[1] F. Benini, E. Barbi, M. Gangemi, L. Manfredini, A. Messeri, and P. Papacci. *Il dolore nel bambino: Strumenti pratici di valutazione e terapia*. Ministero della Salute, 2013 (cit. on pp. 1–3, 6).

[2] *Physiology of pain*. `https://app.lecturio.com/`. Accessed: 2024-02-28 (cit. on p. 2).

[3] G. Zaninetta and S. Presidente. «La Legge 38/2010 e le cure palliative italiane». In: *La voce della SIC* 1 (2010), pp. 8–16 (cit. on p. 2).

[4] «Pain terms: a list with definitions and notes on usage. Recommended by the IASP Subcommittee on Taxonomy». In: *Pain* 6 (1979), pp. 249–252 (cit. on p. 2).

[5] I. A. for the Study of Pain. *Definition of Pain*. `https://www.iasp-pain.org/publications/relief-news/article/definition-pain/`. Accessed: 2024-02-20 (cit. on p. 3).

[6] S. N. Raja et al. «The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises». In: *Pain* 161.9 (2020), pp. 1976–1982. DOI: `10.1097/j.pain.0000000000001939` (cit. on p. 3).

[7] K. D'Apolito. «The neonate's response to pain». In: *MCN: The American Journal of Maternal/Child Nursing* 9.4 (1984), pp. 256–257 (cit. on p. 3).

[8] F. Benini, E. Barbi, and L. Manfredini. «Dolore in Pediatria: miti e verità». In: *Area Pediatrica* 15.4 (2014), pp. 161–172 (cit. on p. 3).

[9] J. Lemons et al. «Prevention and management of pain and stress in the neonate». In: *Pediatrics* 105.2 (2000), pp. 454–461. DOI: `10.1542/peds.105.2.454` (cit. on p. 3).

[10] R. E. Grunau, L. Holsti, and J. W. Peters. «Long-term consequences of pain in human neonates». In: *Seminars in Fetal and Neonatal Medicine*. Vol. 11. 4. Elsevier. 2006, pp. 268–275. DOI: `10.1016/j.siny.2006.02.007` (cit. on p. 3).

[11] R. E. Grunau. «Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity». In: *Rambam Maimonides medical journal* 4.4 (2013). DOI: 10.5041/RMMJ.10132 (cit. on p. 3).

[12] K. Anand and F. M. Scalzo. «Can adverse neonatal experiences alter brain development and subsequent behavior?» In: *Neonatology* 77.2 (2000), pp. 69–82. DOI: 10.1159/000014197 (cit. on p. 3).

[13] B. J. Stevens, R. Pillai Riddell, T. E. Oberlander, and S. Gibbins. «Assessment of pain in neonates and infants». In: *Pain in neonates and infants* 3 (2007), pp. 67–90 (cit. on p. 3).

[14] P. Lago, E. Garetti, G. Boccuzzo, D. Merazzi, A. Pirelli, L. Pieragostini, S. Piga, M. Cuttini, and G. Ancora. «Procedural pain in neonates: the state of the art in the implementation of national guidelines in Italy». In: *Pediatric anesthesia* 23.5 (2013), pp. 407–414. DOI: 10.1111/pan.12107 (cit. on pp. 3, 31).

[15] C. Domenicali, E. Ballardini, G. Garani, C. Borgna-Pignatti, and M. Dondi. «Le scale per la valutazione del dolore neonatale». In: *Medico e Bambino* 33.4 (2014), pp. 223–231 (cit. on pp. 4, 6).

[16] R. V. Grunau and K. D. Craig. «Pain expression in neonates: facial action and cry». In: *Pain* 28.3 (1987), pp. 395–410. ISSN: 0304-3959. DOI: 10.1016/0304-3959(87)90073-X (cit. on pp. 4, 5, 9, 12, 13, 16).

[17] R. Grunau and K. Craig. «Facial activity as a measure of neonatal pain expression». In: *Advances in pain research and therapy* 15 (1990), pp. 147–155 (cit. on pp. 4, 5, 12, 13, 16).

[18] T. Debillon, V. Zupan, N. Ravault, J. Magny, and M. Dehan. «Development and initial validation of the EDIN scale, a new tool for assessing prolonged pain in preterm infants». In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 85.1 (2001), F36–F41. DOI: 10.1136/fn.85.1.F36 (cit. on p. 4).

[19] S. I. Merkel, T. Voepel-Lewis, J. R. Shayevitz, and S. Malviya. «The FLACC: a behavioral scale for scoring postoperative pain in young children». In: *Pediatric Nursing* 23.3 (1997), pp. 293–297 (cit. on pp. 4–6, 13, 33).

[20] J. Lawrence, D. Alcock, P. McGrath, J. Kay, S. B. MacMurray, and C. Dulberg. «The development of a tool to assess neonatal pain.» In: *Neonatal network: NN* 12.6 (Sept. 1993), pp. 59–66 (cit. on p. 4).

[21] R. Carbajal, A. Paupe, E. Hoenn, R. Lenclen, and M. Olivier-Martin. «DAN: une échelle comportementale d'évaluation de la douleur aiguë du nouveau-né». In: *Archives de pédiatrie* 4.7 (1997), pp. 623–628 (cit. on pp. 4, 5).

[22] D. Hudson-Barr, B. Capper-Michel, S. Lambert, T. Mizell Palermo, K. Morbeto, and S. Lombardo. «Validation of the Pain Assessment in Neonates (PAIN) Scale with the Neonatal Infant Pain Scale (NIPS)». In: *Neonatal Network* 21.6 (2002), pp. 15–21. DOI: 10.1891/0730-0832.21.6.15 (cit. on pp. 4, 12, 13, 17).

[23] F. Benini, E. Castagno, and G. P. Milani. «La gestione del dolore nel bambino in pronto soccorso: survey negli ospedali italiani». In: *Quaderni acp* (2019) (cit. on p. 6).

[24] T. M. Heiderich, L. P. Carlini, L. F. Buzuti, R. de C.X. Balda, M. C. Barros, R. Guinsburg, and C. E. Thomaz. «Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units». In: *Jornal de Pediatria* 99.6 (2023), pp. 546–560. ISSN: 0021-7557. DOI: 10.1016/j.jped.2023.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0021755723000669 (cit. on pp. 7, 10, 54).

[25] L. Bergamasco, M. Gavelli, C. Fadda, E. Parodi, C. Bondone, and E. Castagno. «Measurement of Acute Pain in the Pediatric Emergency Department Through Automatic Detection of Behavioral Parameters: A Pilot Study». In: *Bioinformatics and Biomedical Engineering*. Cham: Springer Nature Switzerland, 2023, pp. 469–481. ISBN: 978-3-031-34953-9. DOI: 10.1007/978-3-031-34953-9_37 (cit. on p. 7).

[26] E. Parodi, D. Melis, L. Boulard, M. Gavelli, and E. Baccaglini. «Automated Newborn Pain Assessment Framework Using Computer Vision Techniques». In: *Proceedings of the 4th International Conference on Bioinformatics Research and Applications*. ICBRA '17. Barcelona, Spain: Association for Computing Machinery, 2017, pp. 31–36. ISBN: 9781450353823. DOI: 10.1145/3175587.3175590 (cit. on p. 7).

[27] M. Schiavenato. «Facial expression and pain assessment in the pediatric patient: the primal face of pain». In: *Journal for Specialists in Pediatric Nursing* 13.2 (2008), pp. 89–97. DOI: 10.1111/j.1744-6155.2008.00140.x (cit. on pp. 8–10).

[28] C.-H. Hjortsjö. *Man's face and mimic language*. Studentlitteratur Lund, Sweden, 1970 (cit. on p. 8).

[29] P. Ekman and W. V. Friesen. «Facial action coding system». In: *Environmental Psychology & Nonverbal Behavior* (1978). DOI: 10.1037/t27734-000 (cit. on p. 8).

[30] C. E. Izard. *The Maximally discriminative facial movements coding system (MAX)*. eng. Newark, 1979 (cit. on p. 9).

[31] C. A. Gilbert, C. M. Lilley, K. D. Craig, P. J. McGrath, C. A. Court, S. M. Bennett, and C. J. Montgomery. «Postoperative Pain Expression in Preschool Children: Validation of the Child Facial Coding System». In: *The Clinical Journal of Pain* 15.3 (Sept. 1999), pp. 192–200 (cit. on p. 9).

[32] Y. Sun, C. Shan, T. Tan, X. Long, A. Pourtaherian, S. Zinger, and P. H. de With. «Video-based discomfort detection for infants». In: *Machine Vision and Applications* 30 (2019), pp. 933–944. DOI: 10.1007/s00138-018-0968-1 (cit. on p. 9).

[33] H. Oster. «Emotion in the infant's face: Insights from the study of infants with facial anomalies». In: *Annals of the New York Academy of Sciences* 1000.1 (2003), pp. 197–204. DOI: 10.1196/annals.1280.024 (cit. on p. 9).

[34] X. Cheng, H. Zhu, L. Mei, F. Luo, X. Chen, Y. Zhao, S. Chen, and Y. Pan. «Artificial intelligence based pain assessment technology in clinical application of real-world neonatal blood sampling». In: *Diagnostics* 12.8 (2022), p. 1831. DOI: 10.3390/diagnostics12081831 (cit. on p. 10).

[35] J. Yan et al. «FENP: A Database of Neonatal Facial Expression for Pain Analysis». In: *IEEE Transactions on Affective Computing* 14.1 (2023), pp. 245–254. DOI: 10.1109/TAFFC.2020.3030296 (cit. on pp. 10, 12–14).

[36] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack. «Machine recognition and representation of neonatal facial displays of acute pain». In: *Artificial Intelligence in Medicine* 36.3 (2006), pp. 211–222. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2004.12.003 (cit. on pp. 10, 14, 16, 18).

[37] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack. «SVM Classification of Neonatal Facial Images of Pain». In: *Fuzzy Logic and Applications*. Ed. by I. Bloch, A. Petrosino, and A. G. B. Tettamanzi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 121–128. ISBN: 978-3-540-32530-7. DOI: 10.1007/11676935_15 (cit. on pp. 10, 14, 16, 50).

[38] S. Brahnam, C.-F. Chuang, R. S. Sexton, and F. Y. Shih. «Machine assessment of neonatal facial expressions of acute pain». In: *Decision Support Systems* 43.4 (2007). Special Issue Clusters, pp. 1242–1254. ISSN: 0167-9236. DOI: 10.1016/j.dss.2006.02.004 (cit. on pp. 10, 14, 16, 50).

[39] S. Brahnam, L. Nanni, and R. Sexton. «Introduction to Neonatal Facial Pain Detection Using Common and Advanced Face Classification Techniques». In: *Advanced Computational Intelligence Paradigms in Healthcare – 1*. Ed. by H. Yoshida, A. Jain, A. Ichalkaranje, L. C. Jain, and N. Ichalkaranje. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 225–253. DOI: 10.1007/978-3-540-47527-9_9 (cit. on pp. 10, 14).

[40] S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, and T. Barrier. «Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of Local Descriptors». In: *Applied Computing and Informatics* 19.1/2 (2023), pp. 122–143. DOI: 10.1016/j.aci.2019.05.003 (cit. on pp. 11, 14).

[41] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun. «A Comprehensive and Context-Sensitive Neonatal Pain Assessment Using Computer Vision». In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 28–45. DOI: 10.1109/TAFFC.2019.2926710 (cit. on pp. 12, 14).

[42] P. Hummel, M. Puchalski, S. Creech, and M. Weiss. «Clinical reliability and validity of the N-PASS: neonatal pain, agitation and sedation scale with prolonged pain». In: *Journal of perinatology* 28.1 (2008), pp. 55–60. DOI: 10.1038/sj.jp.7211861 (cit. on pp. 12, 13).

[43] J. Egede, M. Valstar, M. T. Torres, and D. Sharkey. «Automatic Neonatal Pain Estimation: An Acute Pain in Neonates Database». In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 1–7. DOI: 10.1109/ACII.2019.8925480 (cit. on pp. 12, 14, 16, 54).

[44] G. Lu, H. Chen, J. Wei, X. Li, X. Zheng, H. Leng, Y. Lou, and J. Yan. «Video-based neonatal pain expression recognition with cross-stream attention». In: *Multimedia Tools and Applications* 83 (2024), pp. 4667–4690. DOI: 10.1007/s11042-023-15403-z (cit. on pp. 13, 14).

[45] M. S. Salekin, G. Zamzmi, J. Hausmann, D. Goldgof, R. Kasturi, M. Kneusel, T. Ashmeade, T. Ho, and Y. Sun. «Multimodal neonatal procedural and postoperative pain assessment dataset». In: *Data in Brief* 35 (2021), p. 106796. ISSN: 2352-3409. DOI: 10.1016/j.dib.2021.106796 (cit. on pp. 13, 14, 36).

[46] D. Harrison et al. «Too many crying babies: a systematic review of pain management practices during immunizations on YouTube». In: *BMC pediatrics* 14.1 (2014), p. 134. DOI: 10.1186/1471-2431-14-134 (cit. on pp. 13, 14).

[47] D. Harrison, S. Modanloo, A. Desrosiers, L. Poliquin, M. Bueno, J. Reszel, and M. Sampson. «A systematic review of YouTube videos on pain management during newborn blood tests». In: *Journal of Neonatal Nursing* 24.6 (2018), pp. 325–330. ISSN: 1355-1841. DOI: 10.1016/j.jnn.2018.05.004 (cit. on pp. 13, 14).

[48] G. Zamzmi, R. Paul, M. S. Salekin, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun. «Convolutional Neural Networks for Neonatal Pain Assessment». In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1.3 (2019), pp. 192–200. DOI: 10.1109/TBIOM.2019.2918619 (cit. on p. 15).

[49] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang. «Automated assessment of children's postoperative pain using computer vision». In: *Pediatrics* 136.1 (2015), e124–e131. DOI: 10.1542/peds.2015-0029 (cit. on p. 15).

[50] M. Sun, H. Wang, W. Yao, and J. Liu. *AuE-IPA: An AU Engagement Based Infant Pain Assessment Method*. 2022. DOI: 10.48550/arXiv.2212.04764. arXiv: 2212.04764 [cs.LG] (cit. on pp. 15, 18).

[51] L. Celona and L. Manoni. «Neonatal Facial Pain Assessment Combining Hand-Crafted and Deep Features». In: *New Trends in Image Analysis and Processing – ICIAP 2017*. Ed. by S. Battiato, G. M. Farinella, M. Leo, and G. Gallo. Cham: Springer International Publishing, 2017, pp. 197–204. ISBN: 978-3-319-70742-6. DOI: 10.1007/978-3-319-70742-6_19 (cit. on pp. 16, 17, 50).

[52] B. Gholami, W. M. Haddad, and A. R. Tannenbaum. «Relevance Vector Machine Learning for Neonate Pain Intensity Assessment Using Digital Imaging». In: *IEEE Transactions on Biomedical Engineering* 57.6 (2010), pp. 1457–1466. DOI: 10.1109/TBME.2009.2039214 (cit. on p. 16).

[53] T. M. Heiderich, A. T. F. S. Leslie, and R. Guinsburg. «Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements». In: *Acta Paediatrica* 104.2 (2015), e63–e69. DOI: 10.1111/apa.12861 (cit. on pp. 16, 18).

[54] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun. «An approach for automated multimodal analysis of infants' pain». In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016, pp. 4148–4153. DOI: 10.1109/ICPR.2016.7900284 (cit. on p. 17).

[55] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, Y. Sun, and T. Ashmeade. «Automated Pain Assessment in Neonates». In: *Image Analysis*. Ed. by P. Sharma and F. M. Bianchi. Cham: Springer International Publishing, 2017, pp. 350–361. ISBN: 978-3-319-59129-2. DOI: 10.1007/978-3-319-59129-2_30 (cit. on p. 17).

[56] G. Zamzmi, D. Goldgof, R. Kasturi, and Y. Sun. *Neonatal Pain Expression Recognition Using Transfer Learning*. 2018. DOI: 10.48550/arXiv.1807.01631. arXiv: 1807.01631 [cs.CV] (cit. on pp. 17, 38).

[57] G. Zamzmi, R. Paul, D. Goldgof, R. Kasturi, and Y. Sun. «Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN)». In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–7. DOI: 10.1109/IJCNN.2019.8851879 (cit. on pp. 18, 50).

[58] L. P. Carlini, L. A. Ferreira, G. A. S. Coutrin, V. V. Varoto, T. M. Heiderich, R. C. X. Balda, M. C. M. Barros, R. Guinsburg, and C. 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)*. 2021, pp. 394–401. DOI: 10.1109/SIBGRAPI54419.2021.00060 (cit. on pp. 18, 38, 50).

[59] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. *Grad-CAM: Why did you say that?* 2017. DOI: 10.48550/arXiv.1611.07450. arXiv: 1611.07450 [stat.ML] (cit. on pp. 18, 27–29, 46, 47).

[60] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. «A survey of the recent architectures of deep convolutional neural networks». In: *Artificial intelligence review* 53 (2020), pp. 5455–5516. DOI: 10.1007/s10462-020-09825-6 (cit. on pp. 25, 26).

[61] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556 [cs.CV] (cit. on pp. 25, 26, 38).

[62] O. Parkhi, A. Vedaldi, and A. Zisserman. «Deep face recognition». In: *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association. 2015, pp. 1–12. DOI: 10.5244/C.29.41 (cit. on pp. 26, 38, 39).

[63] D. Jin, E. Sergeeva, W.-H. Weng, G. Chauhan, and P. Szolovits. «Explainable deep learning in healthcare: A methodological survey from an attribution view». In: *WIREs Mechanisms of Disease* 14.3 (2022), e1548. DOI: 10.1002/wsbm.1548 (cit. on p. 27).

[64] A. Di Bari, A. Destrebecq, F. Osnaghi, and F. Terzoni. «Traduzione e validazione in italiano della scala Revised FLACC per la valutazione del dolore nel bambino con grave ritardo mentale». In: *Pain Nursing Magazine - ltalian Online Journal* 4-2013 (2013) (cit. on p. 33).

[65] E. Grooby, C. Sitaula, S. Ahani, L. Holsti, A. Malhotra, G. A. Dumont, and F. Marzbanrad. *Neonatal Face and Facial Landmark Detection from Video Recordings*. 2023. DOI: 10.48550/arXiv.2302.04341. arXiv: 2302.04341 [eess.IV] (cit. on pp. 34, 36).

[66] P. Viola and M. 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*. Vol. 1. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517 (cit. on p. 34).

[67] D. E. King. «Dlib-ml: A machine learning toolkit». In: *The Journal of Machine Learning Research* 10 (2009), pp. 1755–1758 (cit. on p. 35).

[68] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. «SSD: Single Shot MultiBox Detector». In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Cham: Springer International Publishing, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2 (cit. on p. 35).

[69] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. «RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 35).

[70] C. Lugaresi et al. «MediaPipe: A Framework for Building Perception Pipelines». In: *CoRR* abs/1906.08172 (2019). arXiv: 1906.08172. URL: http://arxiv.org/abs/1906.08172 (cit. on p. 35).

[71] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann. «Blazeface: Sub-millisecond neural face detection on mobile gpus». In: *arXiv preprint arXiv:1907.05047* (2019). DOI: 10.48550/arXiv.1907.05047 (cit. on p. 35).

[72] G. Jocher. *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*. Nov. 2022. DOI: 10.5281/zenodo.7347926 (cit. on p. 36).

[73] J. Hausmann, M. S. Salekin, G. Zamzmi, D. Goldgof, and Y. Sun. *Robust Neonatal Face Detection in Real-world Clinical Settings*. 2022. DOI: 10.48550/arXiv.2204.00655. arXiv: 2204.00655 [cs.CV] (cit. on pp. 36, 42).

[74] *Keras-VGGFace: VGGFace implementation with Keras framework*. https://github.com/rcmalli/keras-vggface. Accessed: February 21, 2024 (cit. on p. 38).

[75] TensorFlow. *Classification on imbalanced data*. https://www.tensorflow.org/tutorials/structured_data/imbalanced_data. Accessed: 2024-02-23 (cit. on pp. 43, 46).

[76] *StratifiedGroupK-Fold*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html. Accessed: 2024-02-20 (cit. on p. 44).