UNIVERSITY CENTER OF FEI

LUCAS PEREIRA CARLINI

HUMAN VS MACHINE TOWARDS NEONATAL PAIN ASSESSMENT: A COMPARISON OF THE FACIAL FEATURES EXTRACTED BY ADULTS AND CONVOLUTIONAL NEURAL NETWORKS

São Bernardo do Campo

2023

LUCAS PEREIRA CARLINI

HUMAN VS MACHINE TOWARDS NEONATAL PAIN ASSESSMENT: A COMPARISON OF THE FACIAL FEATURES EXTRACTED BY ADULTS AND CONVOLUTIONAL NEURAL NETWORKS

A dissertation submitted for the degree of Master of Science of the University Center of FEI. Supervised by Prof. Dr. Carlos Eduardo Thomaz.

São Bernardo do Campo

2023

Pereira Carlini, Lucas.
Human vs Machine Towards Neonatal Pain Assessment: A Comparison of the Facial Features Extracted by Adults and Convolutional Neural Networks / Lucas Pereira Carlini. São Bernardo do Campo, 2023.
96 p. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Carlos Eduardo Thomaz.

1. Neonatal Pain Assessment. 2. Human Visual Perception. 3. Explainable Artificial Intelligence. I. Thomaz, Carlos Eduardo, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

GENERAL REPORT
MSC DISSERTATION EVALUATION

Postgraduate Research in Electrical Engineering

Master of Science

PGE-10

Name of the candidate: Lucas Pereira Carlini **Register Number:** 121105-1

Title of the Dissertation: Human vs Machine towards neonatal pain assessment: A comparison of the facial features extracted by adults and convolutional neural networks.

Research Area: Signal and Image Processing

Supervisor: Dr. Carlos Eduardo Thomaz

ORIGINAL ASSINADA

Oral Examination Date: 01/27/2023 (mm/dd/yyyy)

Evaluation report (*):

The MSc exam started with the oral presentation of the canditate, which took approximately 30 minutes. After that, both examiners, external and internal ones, individually asked specific and general questions to the candidate about his dissertation and presentation. The candidate properly answered and commented all the points raised by the examiners, showing a consistent knowledge about the research topic of his dissertation and work carried out. In the end, the overall assessment of all examiners is the acceptance of the MSc dissertation.

Please strike out which is not applicable

Final Assessment: (a) ACCEPTANCE ~~(b) REJECTION~~

Based on the overall assessment of all examiners, the examination board shall place the dissertation in one of the two categories above as follows:

(a) Acceptance, indicating in the evaluation report (*) the overall standard attained

(b) Rejection, based on the overall reasons set out in the evaluation report (*)

SIGNATURE OF EXAMINATION BOARD

Dr. Carlos Eduardo Thomaz

Dra. Leila Cristina Carneiro Bergamasco

Dra. Ghada Zamzmi Alzamzmi

Dr. Carlos Eduardo Thomaz
Signature of the Head of PostGraduate Research in Electrical Engineering

To my family, and to the future generation of students and scientists.

ACKNOWLEDGEMENTS

In the middle of 2017, five years ago, Prof. Dr. Carlos Thomaz invited me to his office after I emailed him about an undergraduate research opportunity. At that time, I had no idea that his passion to explain extremely complicated mathematical concepts (yes, I am referring to our conversations about manifold learning) would engage me in such an inspiring way. Since then, we proposed and carried out countless experiments, published several papers, and received many awards and scholarships. But, most importantly, I truly believe that our contributions are positively impacting science and the future generation of researchers. I am very grateful for all his support and mentorship, for all conversations and debates that we had, and, more importantly, for believing in me. I hope that we continue to work for countless years.

Not only that, but I would like to thank all my family. Specially, my gratitude goes to my parents Daniela and Rodrigo; my stepparents Fabio and Karla; my grandparents Zélia, Renato (in memoriam), Elisabeth, and Miguel; and my great-grandmother Carmelina. All of you have supported and helped me whenever I need, and I hope that I was there for all of you as well. In addition, I thank my young siblings Catarina and Dante, since they came to the world, they are my source of happiness.

I would like to name some of my closest friends, Leonardo Ferreira and Gabriel Coutrin. This work would not be possible and would not achieve this quality without the help, contributions, and partnership of both of you. Some parts of the methodology presented here were based on ideas that were sparked on the never-ending discussions (I am to blame for this) that we had. Therefore, I also dedicate this work to you two, and I wish that we continue to work on many more projects. Also, I would like to thank Igor Franco for being one of the best friends that one would desire. I hope that our friendship lasts for life.

Thanks God for being around many lovely, supporting, and caring family and friends.

Finally, I would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and University Center of FEI. The scholarship awarded to me made this work possible.

“All have their worth and each contributes to the worth of the others.”

J.R.R. Tolkien, *The Silmarillion*

“Legacy. What is a legacy?
It’s planting seeds in a garden you never get to see.”

Lin-Manuel Miranda, *Hamilton: The Musical*

ABSTRACT

One of the most important challenges of the scientific community is to mitigate the several consequences for neonates due to pain exposure. This challenge is mainly justified by the fact that neonates are not able to verbally communicate pain, hindering the correct identification of the presence and intensity of this phenomenon. In this context, several clinical scales have been proposed to assess pain, using, among other parameters, the facial features of the neonate. However, a better comprehension of these features is yet required, since some recent results have shown the subjectivity of these scales. Meanwhile, computational frameworks have been implemented to automate neonatal pain assessment. Despite their impressive performances, these frameworks still lack to understand the corresponding decision-making processes. Therefore, we propose to investigate in this dissertation the facial features related to the human and machine neonatal pain assessments, comparing the visual perceived regions by health-professionals experts and parents of neonates with the most relevant ones extracted by eXplainable Artificial Intelligence (XAI) methods using two classification models: (i) VGG-Face, trained originally in facial recognition, and (ii) N-CNN, implemented and trained end-to-end for neonatal pain assessment. Our findings show that the regions used by the classification models are clinically relevant to neonatal pain assessment, yet do not agree with the facial perception of health-professionals and parents. Consequently, these differences suggest that humans and machines can learn with each other in order to improve their current decision-making process of identifying the discriminant information related to neonatal pain. Additionally, we observed that, using the same classification model, the XAI methods implemented here yield distinct relevant facial features to the same input image. These results raise concerns about the effective use and interpretation of XAI methods, and, more importantly, what regions of the image are truly relevant to the decision-making process of the classification model. Nevertheless, our findings advance the current knowledge on how humans and machines code and decode the neonatal facial response to pain. We believe that these findings might enable further improvements in clinical scales and computation tools widely used in real situations, whether based on human or machine decision-making process.

Keywords: Neonatal Pain Assessment. Human Visual Perception. Explainable Artificial Intelligence.

RESUMO

Um dos mais importantes desafios da comunidade científica é mitigar as diversas consequências à exposição da dor em neonatos. Este desafio é principalmente justificado pelo fato de que neonatos não são capazes de verbalizarem dor, dificultando a correta identificação da presença e intensidade deste fenômeno. Neste contexto, diversas escalas clínicas têm sido propostas para avaliar dor, usando, entre outros parâmetros, as características faciais do neonato. Entretanto, uma melhor compreensão dessas caracterísitcas é ainda necessária, visto que resultados recentes demonstraram a subjetividade destas escalas. Enquanto isso, metodologias computacionais têm sido implementadas para automatizar a avaliação da dor neonatal. Apesar de terem desempenho expressivo, estes métodos não permitem a compreensão dos seus processos de tomada de decisão. Portanto, esta dissertação investiga as características faciais relacionadas a avaliação de dor neonatal humana e computacional, comparando as regiões visualmente observadas por profissionais de saúde e pais de neonatos com as características mais relevantes extraídas por métodos de Interpretação de Inteligência Artificial (IIA) utilizando dois modelos de classificação: (i) VGG-Face, treinada originalmente em reconhecimento facial, e (ii) N-CNN, implementada e treinada especificamente para a avaliação de dor neonatal. Os resultados obtidos mostram que as regiões utilizadas pelos modelos de classificação são clinicamente relevantes para a avaliação de dor neonatal, porém, não concordam com a percepção facial de profissionais de sáude e pais. Consequentemente, essas diferenças sugerem que humanos e máquinas podem aprender entre si de forma a melhorar seus atuais processos de identificar as informações discrimimantes da dor neonatal. Adicionalmente, foi observado que, utilizando o mesmo modelo de classificação, métodos distintos de IIA geram diferentes extrações de características para a mesma amostra de imagem de entrada. Estes resultados geram preocupações sobre o efetivo uso e interpretação desses métodos, e, mais importante ainda, quais são as regiões faciais realmente relevantes para o processo de tomada de decisão do modelo computacional. Não obstante, os resultados obtidos avançam o conhecimento atual em como humanos e máquinas codificam e decodificam a resposta facial à dor por neonatos. Estes achados podem permitir o futuro melhoramento das escalas clínicas e ferramentas computacionais utilizadas vastamente em situações reais, sejam esses métodos baseados em decisões humanas ou de máquina.

Palavras-chave: Avaliação de Dor Neonatal. Percepção Visual Humana. Interpretação de Inteligência Artificial.

LIST OF FIGURES

Figure 1 – Hyperplane navigation between “pain” and “no pain” images obtained using the PCA+MLDA framework. 30

Figure 2 – Physical structure built to collect data for the UNIFESP Face Image Dataset. 36

Figure 3 – Examples of images of neonates taken after painful procedure from the UNIFESP Face Image Dataset. 37

Figure 4 – Example of image of a neonate with pain from the iCOPE. 37

Figure 5 – Examples of facial expressions of the iCOPE. 38

Figure 6 – Eye-tracking methodology. 39

Figure 7 – VGG-Face preprocessing. 42

Figure 8 – Illustration of Data Augmentation. 43

Figure 9 – Example of applying Integrated Gradients. 45

Figure 10 – Steps 1 to 3: Integrated Gradients’ attribution mask post-processing. . . . 49

Figure 11 – Step 4: Filtering out less relevant Integrated Gradients’ smoothed attributions using the mean and standard deviation of it. 50

Figure 12 – Example of a contour mask of the most relevant regions according to GradCAM. 51

Figure 13 – Example of a contour mask of the perceived areas by physicians when assessing a neonate with pain. 52

Figure 14 – Samples of input images and GradCAM’s feature extraction on the last convolutional layer of each group of the VGG-Face. 56

Figure 15 – Samples of input images and GradCAM’s feature extraction on the merging and final convolutional layer of the N-CNN when assessing “no pain” (left) and “pain” (right) images. 57

Figure 16 – Samples of heatmaps of facial features extracted by VGG-Face’s (1) GradCAM and (2) Integrated Gradients, N-CNN’s (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts. 59

Figure 17 – Violin plots showing the distributions of the overlap metric to each pair of comparison. 65

LIST OF TABLES

Table 1 – Performance of the VGG-Face architecture. 55

Table 2 – Performance of the adapted N-CNN architecture. 55

Table 3 – Average result of each evaluation metric of all perceptions towards “pain” images. 61

Table 4 – Average result of each evaluation metric of all perceptions towards “no pain” images. 61

Table 5 – p-Value of the statistical difference between “pain” and “no pain” distributions of the comparison of all perceptions. 62

ACRONYMS

AI	Artificial Intelligence
AUC	Area Under the receiver operating characteristic Curve
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
CRIES	Crying, Requires O2 for saturation above 90%, Increased vital signs, Expression, and Sleeplessness
FACS	Facial Action Coding System
GradCAM	Gradient-weighted Class Activation Mapping
iCOPE	infant Classification of Pain Expression
k-NN	k-Nearest Neighbour
LBP	Local Binary Patterns
LSTM	Long Short Term Memory
MLDA	Maximum uncertainty Linear Discriminant Analysis
N-CNN	Neonatal Convolutional Neural Network
NFCS	Neonatal Facial Coding System
NICU	Neonatal Intensive Care Unit
NIPS	Neonatal Infant Pain Scale
NNSOA	Neural Network Simultaneous Optimisation Algorithm
PCA	Principal Component Analysis
PIPP	Premature Infant Pain Profile
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machine
XAI	eXplainable Artificial Intelligence

SYMBOLS

A	feature map of a convolutional layer
α	weight importance of a given feature map
C	number of distinct classes
c	scalar number
ϵ	RMSprop hyperparameter
$F(x)$	function obtained by a classification model
$f(x)$	output given by a activation function in terms of an input sample
γ	interpolation value
i	feature of a given input
k	convolutional layer
L	attribution mask
l_1	weight regularisation
l_2	weight regularisation
M	contour mask
μ	mean of a given set of values
n	number of groups
η	learning rate
Q	concordance indicator
ρ	RMSprop hyperparameter
rg	ranks of a set of values
r_s	Spearman's correlation coefficient
σ	standard deviation of a given set of values
τ	cluster
w	weight between neurons
x	input sample
x'	baseline of an input sample
y	true answer to an input sample
∇	gradient

SUMMARY

1 INTRODUCTION . . . 15
1.1 MOTIVATION . . . 15
1.2 OBJECTIVE . . . 16
1.3 OUTLINE OF THE DISSERTATION . . . 17
2 RELATED WORKS . . . 18
2.1 MEDICAL STUDIES ON NEONATAL PAIN . . . 19
2.1.1 Facial Perception of Pain . . . 26
2.2 COMPUTATIONAL FRAMWEWORKS TO NEONATAL PAIN ASSESSMENT 28
2.3 CURRENT PARADIGM OF RESEARCH . . . 33
3 MATERIALS AND METHODS . . . 35
3.1 FACE IMAGE DATASETS . . . 35
3.1.1 UNIFESP Face Image Dataset . . . 35
3.1.2 infant Classification of Pain Expression . . . 36
3.2 EYE-TRACKING FRAMEWORK . . . 38
3.3 CLASSIFICATION MODELS . . . 40
3.3.1 VGG-Face . . . 40
3.3.2 Neonatal Convolutional Neural Network (N-CNN) . . . 41
3.3.3 Training Protocol . . . 42
3.4 EXPLAINABLE ARTIFICIAL INTELLIGENCE . . . 44
3.4.1 Integrated Gradients . . . 45
3.4.2 Gradient-weighted Class Activation Mapping . . . 47
3.4.3 Attribution Mask Post-Processing . . . 48
3.5 PERCEIVED FACIAL AREAS . . . 51
3.6 IMAGE COMPARISON METRICS . . . 52
4 EXPERIMENTS AND RESULTS . . . 54
4.1 EXPERIMENTS . . . 54
4.1.1 Classification Models Performance . . . 54
4.1.2 Layers' Depth Impact on GradCAM feature extraction . . . 55
4.2 RESULTS . . . 57
4.2.1 Qualitative Results . . . 58
4.2.2 Quantitative Results . . . 60

4.3 DISCUSSION 66
5 CONCLUSION 70
REFERENCES 71
APPENDIX A – FINAL VGG-FACE ARCHITECTURE. 78
APPENDIX B – ADAPTED N-CNN ARCHITECTURE. 80
APPENDIX C – VGG-FACE'S LAYERS DEPTH IMPACT ON GRADCAM FEATURE EXTRACTION ON "PAIN" IMAGES. 82
APPENDIX D – VGG-FACE'S LAYERS DEPTH IMPACT ON GRADCAM FEATURE EXTRACTION ON "NO PAIN" IMAGES. 84
APPENDIX E – N-CNN'S LAYERS DEPTH IMPACT ON GRADCAM FEATURE EXTRACTION ON "NO PAIN" AND "PAIN" IMAGES. . . . 86
APPENDIX F – THE HEATMAPS OF FACIAL FEATURES EXTRACTED BY EACH EXTRACTION MODEL WHEN ASSESSING "PAIN" IMAGES. 88
APPENDIX G – THE PERCEIVED FACIAL AREAS BY EACH EXTRACTION MODEL WHEN ASSESSING "PAIN" IMAGES. 90
APPENDIX H – THE HEATMAPS OF FACIAL FEATURES EXTRACTED BY EACH EXTRACTION MODEL WHEN ASSESSING "NO PAIN" IMAGES. 92
APPENDIX I – THE PERCEIVED FACIAL AREAS BY EACH EXTRACTION MODEL WHEN ASSESSING "NO PAIN" IMAGES. 94

1 INTRODUCTION

Pain is a sensory phenomenon experienced by human beings that acts as an indicator and alert for several unpleasant experiences (DIATCHENKO et al., 2007). This phenomenon might reveal the presence of tissue damage or even a threat to life, triggering degrading physiological and psychological reactions (DIATCHENKO et al., 2007). Also, the constant presence of pain increases the suffering of a certain individual and also results in the reduction of their life span (DIATCHENKO et al., 2006). Consequently, the subject that experiences pain tends to perform an act of escape and withdrawal from the source of this phenomenon (DIATCHENKO et al., 2007). But if pain does occur, the assessment and correct treatment are mandatory for the healthy development of human beings.

Neonates specifically are the primary priority for the continuity of human beings and also for the ongoing evolution of human capital and economic growth (DARMSTADT; SHIFFMAN; LAWN, 2015). Yet, critically ill or preterm babies experience numerous painful procedures during their stay in a Neonatal Intensive Care Unit (NICU) of hospitals (SIMONS et al., 2003; CRUZ; FERNANDES; OLIVEIRA, 2016). These painful experiences can cause degrading effects in the short-, medium-, and long-term in the neonate's life, like, at an early stage of life, irritability and dietary denial (HEIDERICH; GUINSBURG; LESLIE, 2013), but also, it can affect the development of the neonatal brain, leading to disabilities in learning and cognitive tasks (DOESBURG et al., 2013; VINALL et al., 2014; GRUNAU, 2020).

Since neonates are not able to communicate and verbalise pain, several clinical scales have been developed in order to assess pain. These scales mainly use facial expression and its characteristics to recognise the presence of pain (CARLINI et al., 2021b). Although these scales have clinical validity and are used in real situations at NICUs, they are subjective, generating results that may vary according to the characteristics of the patient, the person who performs the assessment, and the environment where both are located (CASSIA XAVIER BALDA; GUINSBURG, 2007). Consequently, the difficulty in identifying and quantifying neonatal pain and the subjectivity of this assessment are an important challenge for the scientific community in general (SOARES et al., 2021).

1.1 MOTIVATION

As mentioned by Salekin et al. (2021a) and Soares et al. (2021), one of the most important challenges of the scientific community is to mitigate the several consequences for neonates due

to pain exposure. Although pain clinical scales have been proposed to identify this phenomenon and use, among other parameters, the assessment of the facial features of neonates (CARLINI et al., 2021b), a better understanding of these features and their discriminative information is yet required, since some recent results have shown the subjectivity of these proposed methods (CARLINI et al., 2020; SOARES et al., 2021; SILVA et al., 2021; BARROS et al., 2021).

In this context, the use of Artificial Intelligence (AI) models allows a non-invasive investigation that is accurate and specific to the pain phenomena felt by neonates. Specifically, eXplainable Artificial Intelligence (XAI) techniques are able to explain the decision-making process that leads an AI model to a particular answer, allowing a better comprehension of such models. This need is also justified by the increasing use of these techniques in medical applications, where reliability and safety are mandatory. Recently, Velden et al. (2021) presented a systemic review of works that used deep learning-based XAI in medical image analysis, categorising their explanation method into three types: visual, textual, and example-based. As shown by the authors, the most common type of medical image analysis is based on visual explanation (VELDEN et al., 2021), presenting frameworks to many anatomical locations of the human body: bladder, brain, breast, cardiovascular, chest, dental, eye, female reproductive system, gastrointestinal, skin, and others. In a different context, Schiller et al. (2020) proposed the analysis of data collected during eye-tracking experiments comparing it with the regions extracted by an AI model when classifying the facial expression of adults (e.g., “neutral”, “happy”, “fear”, etc). This work showed the feasibility of comparing facial features extracted by AI models with the ones observed by human beings.

1.2 OBJECTIVE

The main goal of this dissertation is to investigate the facial features related to the human and machine neonatal pain assessment, comparing the facial information perceived by experts (physicians) and parents of neonates with the most relevant features of two state-of-the-art AI models. The first model was originally trained on facial recognition, whereas the second one was trained end-to-end on neonatal pain assessment. Then, we use XAI methods in order to extract the discriminating regions of neonatal face images, disclosing a better understanding of these often called “black box” models. Specifically, our main research questions are:

a) What are the facial regions most relevant to an AI-based model?

b) Using the same AI-based model, do distinct XAI methods agree with each other?

c) Does an AI-based model agree with the facial perception of Experts and Parents of neonates?

d) Is there any difference in the agreement between sample groups when comparing the perception towards “pain” and “no pain” images?

We believe that with such analysis, we may identify the differences between humans and machines towards their facial processing when assessing neonatal pain, allowing an improvement of the methods currently being used in the clinical practice, and also of the decision-making process performed by AI models.

1.3 OUTLINE OF THE DISSERTATION

This dissertation can be summarised as follows.

In Chapter 2, we describe several works related to neonatal pain of the medical and computational communities. These studies highlight the current clinical understanding of how neonates experience pain, and its consequence on their development and well-being. Also, some of these works investigated how neonates express painful sensations and how adults interpret these responses. Afterwards, we present several computational frameworks to automatically assess neonatal pain.

In Chapter 3, we present our methodology to compare the facial features observed by adults and the ones extracted by AI models to assess neonatal pain. Firstly, we introduce two face image datasets that were used in previous works to develop our framework. Then, we describe our classification models and the XAI methods we used to understand them. Lastly, we detail the perceived facial area extraction, and the image comparison metrics that we used to evaluate the agreement between XAI methods and humans.

In Chapter 4, we show our qualitative and quantitative results, discussing them in order to address our main research questions. Our findings show novel results on the human and machine facial perception when assessing pain, enabling future improvement of clinical scales and computational frameworks.

Finally, in Chapter 5, we conclude this dissertation and discuss the impact of our results regarding the current paradigms of the clinical and computer communities. Additionally, we describe possible future work in order to further investigate the differences between the facial perception of adults and machines.

2 RELATED WORKS

Until the 1970s, there was still no clear definition of human pain. Earlier understanding of pain assumed that this experience could be caused by any cognitive or physical stimulation (OWENS, 1984). It all changed when Merskey (1970) stated that 'Pain is an unpleasant experience which we primarily associate with tissue damage or describe in terms of such damage or both'. However, this definition assumes that the person suffering pain is able to verbally describe such a phenomenon, which infants until a certain age are not capable of doing so. Consequently, physicians and caretakers assumed that infants were not able to feel and suffer pain.

In 1983, Beyer et al. (1983) also observed this assumption when investigating the effective use of analgesics for the relief of pain in adults and children. The results showed that infants, children, and teenagers (between the ages of 1 day and 15 years old) received fewer prescription of potent narcotics than adults (BEYER et al., 1983). In fact, almost all the drugs prescribed to neonates immediately post-operatively were discontinued in the 5th day after surgery, even though the prescriptions were still available. The authors argued that research regarding children's pain tolerance and perception was still ambiguous, since parallel studies had shown that time disagreement about the correlation between pain measure and age (HASLAM, 1969; WOODROW et al., 1972). Additionally, they noted that the reluctance of health professionals regarding overdose, addiction, and respiratory depression often led to doubts about the necessity of analgesic administration. Beyer et al. (1983) was one of the first studies to propose that health professionals must consider the possibility that children may feel pain as intensely as adults, and that the primary difference between the two may be in the accuracy and maturity of their expressions of pain.

This shows that the comprehension of children's pain and, more specifically, neonatal pain and its assessment and treatment is a continuous area of research within the medical scientific community. Although remarkable progress was achieved since the early 1980s, there are concerns that still need to be addressed, specifically about the expressiveness of the neonatal pain and how adults perceive this phenomenon. Meanwhile, studies carried out by computer scientists proposed several frameworks to automatically assess neonatal pain (BRAHNAM et al., 2006, 2007; HEIDERICH; LESLIE; GUINSBURG, 2015; TERUEL et al., 2018; ZAMZMI et al., 2019; SALEKIN et al., 2020, 2021b; CARLINI et al., 2021a). However, except for Heiderich, Leslie, and Guinsburg (2015), these works tackled this problem from either a Medical or Computational

perspective rather than interdisciplinary combining knowledge from the clinical practice to automated models.

In this context, this chapter describes the main works regarding neonatal pain from both scientific communities: medicine and computer science. We close this chapter with our understanding about the current research directions. These works were found on Google Scholar using the following keywords: “neonatal pain assessment”, “neonatal pain facial perception”, “machine learning on neonatal pain assessment”, “automated neonatal pain assessment”, and “explainable artificial intelligence on neonatal pain assessment”. We considered publications with relevant number of citations, regardless of publication date.

2.1 MEDICAL STUDIES ON NEONATAL PAIN

At the end of the 1980s, Anand, Hickey, et al. (1987) provided one of the first literature reviews regarding neonatal pain, covering the functional requirements of neonates for pain perception, the physiologic changes associated with pain, the memory of pain and its behavioural changes. Additionally, the authors proposed the use of the biologic term “nociceptive activity” instead of the subjective term “pain”, since the latter is often related to emotional or cognitive state (ANAND; HICKEY, et al., 1987).

Basically, Anand, Hickey, et al. (1987) described studies that show that neural pathways for pain may be traced from sensory receptors in the skin to sensory areas in the cerebral cortex of newborn infants. These nerves present endings in the skin with similar (or greater) density when compared to endings in adult skin. Beginning with the development of synapses between sensory fibres and interneurons in the dorsal horn of the spinal cord during the sixth week of gestation, the spread of cutaneous receptors starts from the perioral area of the human fetus in the 7th week of gestation, then, it spreads to the hands and feet by the 11th week, to the trunk by the 15th week and all cutaneous and mucous surfaces by the 20th week. On top of that, several researchers suggested in the 1980s that the complete nervous system is active during prenatal development and that detrimental and developmental changes in any part would affect the entire system[1]. Well-defined periods of quiet sleep, active sleep, and wakefulness occur in utero as early as at 28 weeks of gestation, showing that the preterm and full-term babies have various cognitive, coordinating, and associative capabilities in response to visual and auditory stimuli (ANAND; HICKEY, et al., 1987).

[1]To further studies and discussion about the neurophysiology and neurochemical systems related to neonatal pain, see Anand, Hickey, et al. (1987) and Anand and Carr (1989).

Regarding physiologic changes, the pioneering work of Anand, Hickey, et al. (1987) firstly debated over the cardiorespiratory ones. Preterm and full-term neonates during and after a painful procedure (e.g., circumcision or heel lancing) showed increased heart rate and blood pressure. The heart rate variability was related to the intensity and duration of the painful stimulus and to the subjective behaviour of the neonate. Although local anaesthesia prevented these changes in full-term neonates during procedures, the same did not help to alleviate pain in preterm babies. It was also observed that a decrease in transcutaneous partial pressure of oxygen occurred during circumcision, but local analgesic agents prevented these changes. On hormonal and metabolic changes, Anand, Hickey, et al. (1987) stated that preterm and full-term babies who underwent surgery under minimal anaesthesia showed increased release of growth hormone, cortisol, and other corticosteroids. Also, due to suppression of insulin secretion, neonates experienced a severe and prolonged hyperglycaemia. Although the duration was shorter, neonatal stress responses after painful procedure without deep anaesthesia administration were found to be three to five times greater than adults' response. Neonates with potent anaesthetics administration were more clinically stable during surgery and had fewer postoperative complications than neonates under minimal anaesthesia.

Anand, Hickey, et al. (1987) also reviewed studies that investigate the memory of pain in neonates. They reported that, after painful procedure, neonates showed persistence of specific behavioural changes, suggesting the presence of memory. These behavioural changes may, in the short-term, disrupt the adaptation of newborn infants to their postnatal environment, the development of parent-infant bonding, and feeding schedules, while, in the long-term, it could possibly lead to psychological sequelae. Anand, Hickey, et al. (1987) highlighted as well that, even though pain itself cannot be remembered, the experience associated with it can be recalled, as it should, since "painful experience may not be deemed important if it is not remembered".

These behavioural changes can be motor responses, crying, state response (sleeping or awake), and changes in facial expressions. Premature neonates presented lower thresholds for a flexor response and increased sensitisation after repeated stimulation. The reviewed studies reported that full-term babies tend to flex and adduct during the "pinprick in the leg" procedure. Regarding crying, the patterns of it have been correlated with the intensity of pain experienced during circumcision. About the neonate's state, studies observed increased wakefulness and irritability for an hour after circumcision and altered sleep-wake state in neonates undergoing heel-stick procedures. It was also found that, for two days after circumcision, neonates who had received anaesthetics were more attentive to various stimuli and had greater orientation, better

motor responses, decreased irritability, and a greater ability to quiet themselves when disturbed. This suggests that neonates may suffer prolonged effects on the neurological development. Anand, Hickey, et al. (1987) also depicted some studies on the facial expression related to pain. Their initial findings showed that distinct neonatal facial expressions can be discriminative in pain assessment as well, and may be altered by the behavioural state of the neonate.

In 1999, Guinsburg (1999) estimated that a neonate in a NICU receives from 50 to 150 painful procedure per day. Regarding neonates under 1 kg, this number can increase to over 500 painful procedures. Specifically on this topic, Southall et al. (1993) observed earlier that, in a cohort of 55 patients, 181 painful procedures were carried out, and 50 of them had been done without analgesia or sedation. Southall et al. (1993) also noted that the longer the stay, more painful procedures were performed . These results highlight the need for reliable pain assessment tools to be used in clinical situations.

To this end, Grunau introduced the Neonatal Facial Coding System (NFCS) (GRUNAU; CRAIG, 1987). The proposed system was based on the Facial Action Coding System (FACS) approach (P.; FRIESEN, 1978) with adaptions for coding babies' faces (OSTER, 1978). Using the NFCS, participants of the experiment carried out by Grunau identified that, following heel-lance which is an invasive and painful procedure, neonates early in the second day of life showed a number of facial changes, namely eye squeeze, brow contraction, nasolabial furrow, taut tongue, and open mouth, that differed substantially from the amount and type of facial action provoked by heel-rub, a distressful but not painful procedure. Also, Grunau and Craig (1987) observed that the current behavioural state (sleep or awake) of the neonate affects its pain expression. Infants quietly sleeping showed the least facial reaction and the longest latency to cry. Overall,this pioneering work was the first to establish specifically that facial features are a reliable and useful measure of pain expression (GRUNAU; CRAIG, 1987).

Few years later, Grunau, Johnston, and Craig (1990) carried out a new experiment with recently born neonates (mean age of 2.2 hours) when facing: an injection of vitamin K (invasive and painful procedure), application of a disinfectant solution to the stub of the umbilical cord, and rubbing the thigh with an alcohol swab (both non-invasive and distressful). Similarly to the previous study, volunteers using the NFCS identified that specific facial actions, such as brow bulging, eyes squeezed shut, deepening of the nasolabial furrow, and open mouth, were associated more frequently with the invasive procedure. Additionally, some neonates did not cry during painful procedures, but it occurred in response to non-painful events, raising doubt about this behavioural change as a reliable and solely source to identify pain.

Subsequently, Craig et al. (1994) investigated the correlation between the NFCS and FACS. Since the former is a subset of facial features of the latter, the authors were concerned if the NFCS would yield an incorrect interpretation of the infant expression of pain that would be better described using FACS. The results demonstrated that the NFCS captures correlated features to the FACS. However, the NFCS presents significant advantages over the FACS due to its simpler and less complex set of features to be learned and mastered. Also, these results highlight the fact that the facial expression of pain of adults and neonates is similar, consequently, endorsing the argument that infants are born with the capacity to experience and express pain (CRAIG et al., 1994).

Later on, Guinsburg (1999) reviewed the other following neonatal pain scales that were proposed and validated in clinical situations: Neonatal Infant Pain Scale (NIPS) (LAWRENCE et al., 1993), Postoperative Pain Measurement Score (CRIES – Crying, Requires O2 for saturation above 90%, Increased vital signs, Expression, and Sleeplessness) (KRECHEL; BILDNER, 1995), and the Premature Infant Pain Profile (PIPP) (STEVENS et al., 1996). Differently from the NFCS, these are multimodal scales, including the analysis of the facial expression, cry, body motion, behavioural state and physiological signs.

In more details, the NIPS is composed of the observation of the presence of cry (1 point: whimper; 2 points vigorous cry), change of breathing (1 point: irregular, gagging or breath holding), flexed/extended arms or legs (1 point), fussy state of arousal (1 point), and facial expression (1 point: grimace). Pain is present if the total score is over 3 (LAWRENCE et al., 1993). The NIPS is recommended to assess acute procedural pain in infants less than 1-year-old.

As its name suggests, CRIES is based on the analysis of the crying, oxygen saturation, vital signs, facial expression and sleeplessness (KRECHEL; BILDNER, 1995). Each feature has a score that ranges from 0 to 2. Experimental results showed that the assessment of pain had a median score of 4. It should be noted that, differently from the other scales, CRIES is meant to be used in postoperative monitoring of neonates not intubated and paralysed.

The last pain scale investigated (GUINSBURG, 1999) PIPP assesses acute pain in preterm and term neonates (STEVENS et al., 1996). Fewer gestational weeks lead to higher score for each of the analysed characteristics. The caretaker perceives the neonate for 15 seconds, annotating the heart rate and oxygen saturation, and the behavioural state (quiet/active, sleep/awake, facial movements and crying). Then, for the next 30 seconds, the caretaker must distinguish if the neonate presents an increase in heart rate and a decrease in the oxygen saturation. Also, the

caretaker must observe the neonate's brow, eye and nasolabial furrow. For a more detailed description of this scoring, see Stevens et al. (1996).

In 2006, Hummel and Dijk analysed the challenges in behavioural pain assessment, given the enormous progress achieved in the 1990s and early 2000s (HUMMEL; DIJK, 2006). According to the authors, behavioural and even physiological changes may occur during non-painful, but distressful scenarios. The difference between these two behavioural states is difficult to identify, since painful situations are also distressful, but not all experiences of distress are pain related, such as hunger, maternal deprivation and ventilatory factors. The authors also observed that sedative administration and neuromuscular blocking agents can mitigate or inhibit the behavioural expression of pain (HUMMEL; DIJK, 2006). In fact, all the aforementioned pain scales are based on the behavioural changes in terms of a normal baseline, consequently, their use may be unreliable in such situations. Likewise, severity of illness might also decrease the infant's ability to exhibit pain signals (HUMMEL; DIJK, 2006). According to the authors, preterm neonates are unable to communicate pain behaviourally as readily as term infants, and may present decreased behavioural and increased physiological response to heel-stick after recurrent painful procedures (HUMMEL; DIJK, 2006). Specifically analysing multimodal scales that include behavioural and physiological pain assessment, Hummel and Dijk (2006) referenced adult studies which suggest that there is no correlation between both indicators. Even when adults are behaviourally stable due to sedation, physiological reactivity to tissue damage still occurs. More broadly, both indicators might not be correlated across distinct age groups or patients with different diseases (HUMMEL; DIJK, 2006). It is noteworthy that the behavioural and physiological features of the CRIES and PIPP present the same relevance to its composite score. Another equally important challenge is the assessment of ongoing or chronic pain (HUMMEL; DIJK, 2006). The scales previously described aim to assess characteristically acute procedural pain. The identification of pain expression during these situations is complicated by non-specific behavioural response, turning coding systems unfeasible. Therefore, the clinical, but subjective, judgment is the sole source to determine the causality of the behaviour (HUMMEL; DIJK, 2006).

Even though there were concerns about acute, ongoing/prolonged, and chronic pain, its clear definitions to neonates were not established until recently. In 2017, Anand (2017) stated that few methods to assess the intensity of prolonged and chronic pain were proposed, but due to the absence of precise definitions, other aspects related to chronic pain have not been dealt with, such as duration, periodicity and secondary effects. Furthermore, Anand (2017) reviewed that only 10% of neonates in NICU received daily clinical assessments for prolonged and ongoing

pain. Anand's work suggests that this is due to scarce clinical examples of ongoing pain. The author also states that most clinicians can easily assess the presence of prolonged or chronic pain in specific situations, but there is still a need for developing the taxonomy of pain terms dedicated for neonates. To this end, Anand (2017) proposed, as a starting point for defining pain in newborns, the following features: temporal, character of pain, secondary effects and response patterns.

According to Anand (2017), the beginning and duration define any painful experience, therefore, differentiating acute and non-acute pain. Acute pain occurs with the onset of tissue damage, and it usually lasts during the painful stimulus (or for brief periods thereafter). Meanwhile, chronic pain depends on further characteristics, such as the length of the neonatal period, development of long-term effects of pain, and tolerance to analgesia. With this in mind, Anand (2017) stated that pain lasting longer than a week shall be defined as chronic pain in neonates. Prolonged or persistent pain duration ranges from one hour to seven days.

Regarding the character of pain, there are two features specific to the developing nervous system (ANAND, 2017). Firstly, the immaturity of the central and peripheral nervous systems has lower activation, excitation, and transmission thresholds to nociceptive stimuli when compared to older ages. Also, stimulation of the receptive field in the spinal cord heightens nociceptive signalling and can evoke a long-lasting excitability within it. Only during infancy, the inhibitory signalling in the spinal cord is developed. These features make clinicians attempts to localise pain in neonates to result in failure, while increasing the chances of secondary effects. As such, acute pain is sharp and well-localised, and prolonged and chronic pain is majorly dull and diffusely localised (ANAND, 2017).

Tissue damage or inflammation causes secondary effects, such as hyperalgesia and allodynia (ANAND, 2017). Hyperalgesia is the increased pain to a stimulus that is normally painful, and allodynia is the pain due to stimuli that rarely provoke pain. Anand (2017) reviewed some studies that suggest hyperalgesia after heel lances, venipuncture and abdominal surgery, but further information is still scarce. With respect to allodynia, some studies show that preterm neonates appear to suffer it, presenting similar responses to painful and non-painful stimulus (HOLSTI et al., 2005; SLATER et al., 2010; FABRIZI et al., 2011; CORNELISSEN et al., 2013). Anand (2017) declares that standardised tests for allodynia need to be developed and validated.

Anand (2017), also observed that behavioural and physiological responses are well-defined for acute pain. However, studies related to prolonged or chronic pain presented assessment

methods with overlapping features and different from acute pain. Some physiological and behavioural responses may not occur during prolonged or chronic pain.

More recently, Williams and Lascelles (2020) reviewed the current comprehension of the consequences of pain exposure at early life. Mainly, these situations occur due to diagnostic procedures (heel stick, venipuncture, retinopathy of prematurity exam, etc), therapeutic procedures (chest physiotherapy, tracheal intubation/extubation, etc) and surgical procedures (circumcision, cardiac surgery, etc). The number of procedures is also correlated with the length of stay in the NICU. And it is more concerning to preterm babies, since the average length of stay is 81 days, while term neonates present an average length of 5 days. Additionally, Williams and Lascelles (2020) observed that the demographics of the NICU are predominately composed by premature newborns (72.3%) when compared to term newborns (27.7%). Yet, Williams and Lascelles (2020) reviewed studies that show that still most neonates receive none (or inadequate level of) analgesia. These factors can cause negative long-term effects on neonates related to cognitive impairments, psychosocial dysfunction, increased pain sensitivity and poor health outcomes.

As shown by Williams and Lascelles (2020) as well, painful procedures in early life are associated with a brain response reduction to non-painful stimulus. Prematurity seems to aggravate this deficiency too. Furthermore, these procedures can negatively affect brain growth, resulting in lower intelligence, language and attention deficits, poor visual-motor functions, and poor behavioural outcomes. For preterm neonates, these outcomes may persist into adolescence and young adulthood. Males appear to be at a greater risk than female preterm babies (WILLIAMS; LASCELLES, 2020).

Regarding psychosocial dysfunction, Williams and Lascelles (2020) additionally analysed that early repetitive pain may cause attention deficit disorders and atypical behaviours, such as hypervigilance and exaggerated startled responses. Neonates may suffer from internalising behaviours, leading to depression and anxiety and, ultimately, greater social difficulties. Preterm neonates have a greater chance of presenting internalising behaviours.

Williams and Lascelles (2020) reported also that early pain is associated with hypersensitivity later in life. This effect is increased in extremely preterm neonates who underwent prior surgery. Following secondary painful stimuli, neonates with a history of pain may present increased heart rates and lower oxygen saturation (WILLIAMS; LASCELLES, 2020). This work investigated subsequent poor health outcomes. Increased chances of diabetes, hypertension and high blood pressure later in life are associated with preterm neonates which experienced early

pain, exposing them to cardiovascular diseases and stroke. Moreover, adults with these health conditions are more likely to suffer from chronic pain.

2.1.1 Facial Perception of Pain

Since facial features are assessed in the majority of pain scales used by health professionals during clinical training and practice, we complement this clinical review with works that specifically investigated the facial perception of adults when assessing neonatal pain.

One of the first works on the adults' facial perception of the neonatal response to pain is from Balda et al. (2000). Specifically, the authors investigated whether health professionals and parents could recognise the "pain" image within a set of photos of neonates. Each adult observed 3 sets of 8 images (1 photo taken after heel puncture and the remaining ones during rest, light stimulus and heel friction). The authors screened all pictures with the NFCS and only the photos shot after heel puncture scored over 3 were classified as "pain". Regarding the first set of pictures, health- and non-health professionals presented similar accuracy to identify the correct photo of pain. However, non-health professionals were more accurate in recognising the photo of neonatal with pain in the second and third set. The authors suggested that health professionals, with more theoretical knowledge and professional experience, may underestimate the pain of their patients and, consequently, modify their facial perception when assessing neonatal pain (BALDA et al., 2000).

More recently, we proposed a novel computational framework using eye-tracking technology to investigate the human facial perception related to neonatal pain assessment (CARLINI et al., 2020). Our experiments collected visual data from 44 physicians, 40 healthcare professionals, 29 parents, and 30 laypersons when observing 20 images of neonates during rest and after a painful procedure. Additionally, for each assessment, each participant verbally assigned a pain score, ranging from 0 (no pain) to 10 (extreme pain). To categorise these evaluations, numerical scores greater than or equal to 3 were classified as "newborn with pain" and scores less than 3 as "newborn without pain" (CARLINI et al., 2020). We carried out all procedures at the Hospital São Paulo of the Escola Paulista de Medicina of the Universidade Federal de São Paulo.

These preliminary results showed that all groups of volunteers have a disagreement regarding the intensity of pain, with a high variance of the assigned scores (CARLINI et al., 2020). For images of neonates at rest, the scores were concentrated close to zero. When comparing distinct sample groups, physicians and health professionals showed higher accuracy than parents

and laypeople to correctly identify the absence of pain and slightly better to identify its presence. Holistically analysing the facial perception of these groups, there was a general interest in the regions of eyes, nose, and mouth. Specifically analysing the areas of the already described NFCS (GRUNAU; CRAIG, 1987; GRUNAU; JOHNSTON; CRAIG, 1990), no apparent difference was identified in the facial perception of these adults. In other words, physicians and health professionals with clinical training showed similar holistic facial perception to adults without clinical practice (CARLINI et al., 2020).

Subsequently, Soares et al. (2021) proposed a feature-by-feature analysis to discriminate the gaze of health professionals and non-professionals. Both groups of volunteers showed holistic facial perception, but with subtle differences. Health professionals looked more at the regions of the mouth and nasolabial furrow and less at the eyes and forehead. Moreover, the correlation of gaze between these groups was almost perfect given the number and time of fixations for all the facial regions analysed, except for the number of fixations in the nasolabial furrow (SOARES et al., 2021). These results highlight and agree with our findings (CARLINI et al., 2020), indicating the homogeneity of adults' facial perception when assessing pain in neonates, regardless of clinical training or experience.

Silva et al. (2021) investigated the differences in physicians' gaze when classifying as well pain and non-pain images of neonates. The authors identified that physicians looked for a greater number of times at the mouth, eyes, forehead, than to the nasolabial furrow when analysing pain images. Regarding the time of fixation at each facial feature, physicians looked for a longer period at the mouth, eyes, forehead, and less at the nasolabial furrow when observing images of neonates after a painful procedure. Analysing the time of fixation in each facial feature, each additional second of gaze in the mouth and forehead increased the chance of correctly identifying moderate or severe pain (SILVA et al., 2021). Silva et al. (2021) hypothesised that the gaze in these regions is related to the detection of discriminant signals, with an unconscious emotional response on an affective-cognitive level as a consequence of past experiences.

Differently from previous works, Barros et al. (2021) proposed the use of a latent class analysis technique (DE LA TORRE, 2011) to investigate differences in the adult facial perception. In this way, all volunteers of the experiments (CARLINI et al., 2020) were divided into: adults who correctly identified the presence and absence of pain (class YY), those who correctly classified only the presence of pain (YN), those who correctly classified only the absence of pain (NY), and, finally, those who did not correctly identify any newborn image (NN). The evaluations of images of neonates in pain were considered correct when the adult attributed a

score greater than or equal to 6. For the images of neonates at rest, only the scores less than or equal to 2 were considered correct (BARROS et al., 2021). Similar to the previous results, all classes presented a higher concentration of gaze on the forehead, eyes, and mouth, rather than on the nasolabial furrow. However, adults who correctly identified the presence and absence of pain (class YY) presented a higher number and time of fixation at the nasolabial furrow. In other words, each additional fixation in this region resulted in a greater chance of belonging to class YY. The authors also identified that being a health professional increases the probability of being class YY (BARROS et al., 2021).

We presented very recently a systematic review of several clinical scales for the classification of pain in newborns (CARLINI et al., 2021b; TAMANAKA et al., 2022). In particular, we proposed a comparative analysis of the importance of each facial region for these same scales and for adult facial perception (TAMANAKA et al., 2022). Using the Kendall correlation (KENDALL, 1938), we identified that the facial features proposed by the literature do not have the same relevance as the regions observed by adults, such as the nose and cheeks. Furthermore, when applying this same correlation in a comparison between groups of participants, physicians, and health professionals showed a higher correlation with each other than with parents or laypeople (TAMANAKA et al., 2022).

2.2 COMPUTATIONAL FRAMWEWORKS TO NEONATAL PAIN ASSESSMENT

To the best of our knowledge, Brahnam et al. (2006) was the first work to approach neonatal pain assessment through a machine learning framework. The authors created the iCOPE dataset, used here and fully described in Section 3.1.2, to perform their experiments, labelling the photos taken during painful procedure as “pain”, and the pictures taken during air puff, crying, friction of the heel and resting as “no pain”. The proposed framework comprehends a preprocessing stage, a feature extraction one, and, then, the classification.

During preprocessing, the authors converted each image to grey scale, rotated, cropped and scaled, aligning the eyes along the same axis. Additionally, the region of the face was centred within an ellipse. The pixels within it resulted in a feature vector. The authors applied Principal Component Analysis (PCA), extracting the most relevant features of the obtained vector dataset (JOLLIFFE; SPRINGER-VERLAG, 2002; FUKUNAGA, 2013). Support Vector Machine (SVM) (BOSER; GUYON; VAPNIK, n.d.; CRISTIANINI; SHAWE-TAYLOR, et al., 2000) classified the input as “pain” or “no pain”.

Brahnam et al. (2006) carried out several tests with different kernels to the SVM. Their experimental results ranged from 70.00% accuracy with the Radial Basis Function (RBF) kernel to 88.00% with a polynomial kernel of third degree. Moreover, PCA showed to be an important step of the framework, increasing the accuracy from 85.75% to 88.00% and reducing the testing time per image from 0.00780s to 0.0008s to the SVM with a polynomial kernel of third degree. The authors did not apply any statistical significance test in the results.

Following this pioneering work, Brahnam et al. (2007) proposed a more realistic training protocol, named leave-one-subject-out, in which training sets did not include the subject of the test set. Furthermore, the authors applied for the first time the Neural Network Simultaneous Optimisation Algorithm (NNSOA) (SEXTON; DORSEY; SIKANDER, 2004) on a medical task.

The obtained results showed that the NNSOA outperformed the SVM, regardless of the kernel. SVM with a linear kernel provided the best SVM classification rate of 82.35% accuracy, while the NNSOA achieved 90.20% accuracy. Admittedly, these results did not present statistical difference between them. However, as shown by the authors, NNSOA seemed to be more stable, since it had the lowest standard deviation.

In both studies, Brahnam and collaborators did not investigate the performance of the studied models using the F1 score (SASAKI et al., 2007) and the Area Under the receiver operating characteristic Curve (AUC). These metrics are important for two main reasons: (1) the iCOPE is not a balanced dataset, meaning that the number of "pain" samples is not equal to "no pain", and (2) it provides a better understanding of the generalisation and class-distinction capabilities of the models, since it account precision, recall and false positive rate. These capabilities are relevant to clinical situations because over- and under-treatment of pain can lead to several consequences for the patient.

Almost a decade later, Heiderich, Leslie, and Guinsburg (2015) proposed a computational framework based on the facial features of NFCS (GRUNAU; CRAIG, 1987; GRUNAU; JOHNSTON; CRAIG, 1990). To this end, the authors also created the UNIFESP Face Image Dataset, used here as well and fully described in Section 3.1.1. For each image of the dataset, the authors identified 16 facial landmarks, related to the forehead, nasolabial furrow, and mouth, and calculated the Euclidean distance between each one of them.

To discriminate "pain" from "no pain" images, the authors created a standard "no pain" set of distances based on the photos of neonates during periods of rest. Similarly to the NFCS (GRUNAU; CRAIG, 1987; GRUNAU; JOHNSTON; CRAIG, 1990), when the neonate of the test image presented more than 3 facial features, this image was classified as "pain". The proposed

system achieved 85% sensibility and 100% specificity to “no pain” images and 100% sensibility and specificity to “pain” images. Additionally, it should be noted that the authors implemented the framework in hardware to be used in real clinical situations.

In 2018, Teruel et al. (2018) applied a statistical pattern recognition framework on the UNIFESP Face Image Dataset. The framework, originally proposed by Thomaz et al. (2007), enables a feature extraction analysis, allowing a better understanding of the features associated with “pain” and “no pain” images, and it is composed of two main steps: (1) dimensionality reduction using PCA (JOLLIFFE; SPRINGER-VERLAG, 2002; FUKUNAGA, 2013), and (2) feature extraction and classification using the Maximum uncertainty Linear Discriminant Analysis (MLDA) (FISHER, 1936; THOMAZ; KITANI; GILLIES, 2006). The last method calculates the most discriminative hyperplane between each class of image. Therefore, it was possible to navigate within each region and, when reconstructing the sample to its original dimension (image size), to visualise the distinctions between classes.

Analysing the results, images classified as “pain” seem to be associated with the open mouth and deep nasolabial furrow, showing agreement with the NFCS. Quantitatively, the mean accuracy of the framework achieved 72.77%.

Figure 1 – Hyperplane navigation between “pain” and “no pain” images obtained using the PCA+MLDA framework.

Source: Adapted from Teruel and Thomaz (2018).

More recently, Zamzmi and research collaborators achieved impressive results using Convolution Neural Networks (CNNs) in neonatal pain classification (ZAMZMI et al., 2018, 2019; SALEKIN et al., 2021c).

Firstly, in Zamzmi et al. (2018), the authors proposed the Transfer Learning technique in order to use a pre-trained CNN in a different task from the original one. The authors used 4 different CNNs architectures: 3 of them (VGG-F, VGG-M and VGG-S, proposed by Chatfield et al. (2014)) were trained in object recognition, and the last one (VGG-Face, proposed by Parkhi, Vedaldi, and Zisserman (2015)) was trained in facial recognition. Furthermore, the authors created a new dataset with neonatal images classified as “pain” or “no pain” using NIPS (LAWRENCE et al., 1993). The ZFace face detector algorithm extracted the corresponding

face of each image of the dataset (JENI; COHN; KANADE, 2015). Then, the authors reshaped the obtained images to 224x224, following the CNNs' input size. For each CNN architecture, the authors extracted the features from the last fully connected layer before the output layer (high-level and more relevant to the original CNN's task), and from the last convolutional layer (low-level and generic features). The authors selected the most relevant features (5, 10, and 15 features) using the Relief-f and Symmetric Uncertainty (KIRA; RENDELL, 1992; HALL, 1999). The obtained samples were classified using Naive Bayes, k-Nearest Neighbours (k-NN), SVMs, and Random Forests (RFs), as they had shown surprising performance with Transfer Learning (BAR et al., 2015; PAUL et al., 2016; RASSADIN; GRUZDEV; SAVCHENKO, 2017; SARGANO et al., 2017). Regarding the VGG-F, M, S architectures, VGG-S achieved the best results with high-level features (90.41% accuracy and 0.742 AUC) and the VGG-F achieved the best result with low-level features (87.13% accuracy and 0.713 AUC). Even though the accuracy using high-level features is higher, the AUC difference was not statistically significant. To the VGG-Face architecture, the use of high-level features achieved 90.34% accuracy (2.3% higher than low-level features) and 0.841 AUC (not statistically significant).

Then, in Zamzmi et al. (2019), the authors proposed the first CNN architecture specifically trained to neonatal pain assessment. The Neonatal Convolutional Neural Network (N-CNN) is a cascaded CNN composed of three branches followed by convolutional, pooling and fully connected layers. The first branch is composed of a single max pooling layer, capturing image-specific information, whereas the second and third branches are composed of convolutional and pooling layers, extracting generic information, such as edges and blobs. The features obtained in each branch are merged and processed in another pair of convolutiona-pooling layer. A fully connected layer and sigmoid layer performs the classification. The authors stated that they carried out several experiments using different regular and cascades architectures, but the selected one achieved the best results (ZAMZMI et al., 2019). Using a private dataset (ZAMZMI et al., 2019) and the iCOPE (Section 3.1.2), the authors compared the N-CNN with the ResNet architecture and with a classifier based on Local Binary Patterns (LBPs) and SVMs. Regarding the first dataset, the N-CNN achieved 91% accuracy and 0.93 AUC, while the ResNet reached 87.1% accuracy and 0.89 AUC, and the LBP obtained 85.5% accuracy and 0.82 AUC. To the iCOPE, the N-CNN, ResNet and LBP achieved, respectively, 84.5%, 82.9% and 81.3% accuracy, without statistical significance tests.

Salekin et al. (2021b) presented recently a temporal multimodal deep learning approach to specifically assess neonatal postoperative pain. According to the authors, the assessment of

postoperative pain requires a continuous and frequent observation of the neonate due to the ever-changing intensity and expression of pain. Additionally, the authors argue that a multimodal approach is necessary to an AI-based model, since several clinical validated assessment tools are based on at least two dimensions. In this context, the authors proposed the analysis of the facial expression combined with the body movement and crying sound. The final classification is obtained by a majority voting of the labels of each indicator. The facial expression analysis is based on a Bilinear CNN with two VGG-16 architecture (PARKHI; VEDALDI; ZISSERMAN, 2015), followed by two fully connected layers and a single dense layer, and a Long Short Term Memory (LSTM) architecture of Recurrent Neural Network (RNN) in order to integrate the temporal information. The Bilinear CNN is used to extract features from two different regions of the input image, enabling fine-grained image classification (LIN; ROYCHOWDHURY; MAJI, 2015). Meanwhile, the LSTM architecture learns the temporal information of the deep features extracted by both VGG-16 CNNs (HOCHREITER; SCHMIDHUBER, 1997). To body movement feature extraction, Salekin et al. (2021b) trained the VGG-16 using a motion and body image of the neonate. The motion image is calculated by subtracting consecutive frames of the video and attributing a value of 1 (movement) or 0 (no movement). In this way, the motion image segments regions of the input sample that presents movement. Similarly to the facial expression analysis, the deep features of body movement obtained by the VGG-16 are processed on the same, but parallel to the facial expression one, LSTM architecture. Regarding the crying sound analysis, the authors converted the raw audio signal to a spectrogram image, obtaining a visual representation of the input audio signal. Then, the VGG-16 extracted the deep features of the image. According to the authors, the spectrogram image represents the change of frequency components with respect to time and suppresses noise.

Observing the results obtained by Salekin et al. (2021b), the crying sound presented higher relevance than the analysis of the face and body movement when assessing postoperative pain. It achieved 79.63% accuracy and 0.87 AUC, outperforming the 69.52% accuracy and 0.82 AUC of the face and the 70.50% accuracy and 0.78 AUC of the body. The authors suggest that the crying sound can better assess postoperative pain as compared to facial expression and body movement (SALEKIN et al., 2021b). With respect to the multimodal approach, combining the labels of each indicator, it achieved similar results to the crying sound analysis. The only observed difference is the AUC, in which the multimodal approach achieved 0.90 performance, without any statistical significance test to compare the performance of each approach (multimodal against unimodal analysis). The authors state that, even though the crying sound presented similar

performance to the multimodal approach, the latter is still necessary because pain manifests itself in different signals and there might be situations that one indicator is unavailable (SALEKIN et al., 2021b).

In 2021, we implemented a deep learning-based approach to be used on mobile devices (CARLINI et al., 2021a). Using the UNIFESP and iCOPE datasets, we fine-tuned the VGG-Face architecture (PARKHI; VEDALDI; ZISSERMAN, 2015) and trained from scratch two fully connected layers with 512 neurons each. This architecture was found by a random search with parameters ranging from 50 to 2048 neurons and 1 to 3 fully connected layers. Additionally, we observed the influence of training the model with each and both datasets. The evaluation process was carried out using the hold-out cross-validation technique (10 repetitions), randomly choosing independent test sets (20% of original datasets) for each repetition. When trained with the UNIFESP dataset only, the model presented 72% accuracy and 0.74 AUC. The iCOPE-trained model outperformed it, achieving 83% accuracy and 0.82 AUC. However, as expected, the model trained with both datasets obtained the best performance with 89% accuracy and 0.86 AUC. The differences between evaluation metrics were statistically significant at the $p = 0.05$ level. The best model of the hold-out repetitions was later optimised for mobile environment, and embedded in an application developed for Android devices (CARLINI et al., 2021a). In this work, we also described results when applying XAI techniques in neonatal pain assessment. Using the Integrated Gradients algorithm, which will be described in Section 3.4.1, we showed that the classification model highlighted regions of the face similar to the facial perception of adults when assessing pain and similar to the regions proposed by clinical scales, named the nasolabial furrow, forehead, mouth, and protrusion of the tongue.

2.3 CURRENT PARADIGM OF RESEARCH

We have described here several studies proposing clinical scales to assess pain through the behavioural and physiological response of the neonate. Many of these studies also suggested that these responses may be impacted by the current state of the neonate (e.g., awake, sleeping, sedated, etc), the severity of illness, and the frequency of painful procedures/situations. Also, these responses may occur during distressful, but not painful situations. More recently, research suggests that the human facial perception when assessing neonatal pain is an innate cognitive task that, holistically, is independent of training or clinical experience, even though some subtle differences between health and non-health professionals might be identified. The fact that facial

features proposed by the related literature and clinical scales do not have the same relevance as those that compose the human facial perception is intriguing. Overall, we might say that these works show that clinical scales might need to be reviewed or revisited, considering non-subjective and pain-specific information.

One of the possible approaches to address these issues is the use of computational frameworks specifically implemented end-to-end to pain assessment. Even though the studies described here showed impressive performance to discriminate “pain” inputs from “no pain” ones, there is still a lack of pain-related works proposing to understand the decision-making process of the classification models, which can be achieved through the use of XAI methods. These investigations can lead, ultimately, to a better non-subjective and pain-specific findings of the several behavioural and physiological responses of the neonate, enhancing the human knowledge towards neonatal pain assessment.

3 MATERIALS AND METHODS

In this chapter, we present two face image datasets containing several photos of neonates during periods of rest (labelled as “no pain”) and during painful procedures (“pain”). Then, we follow with the eye-tracking framework implemented (CARLINI et al., 2020). Also, we detail the classification models, and the XAI methods that we used to investigate them. Subsequently, we describe the post-processing steps to extract the perceived facial areas by humans and machine. To the end of this chapter, we describe the evaluation metrics used to compare the human facial perception with the most relevant regions of the image to each model.

3.1 FACE IMAGE DATASETS

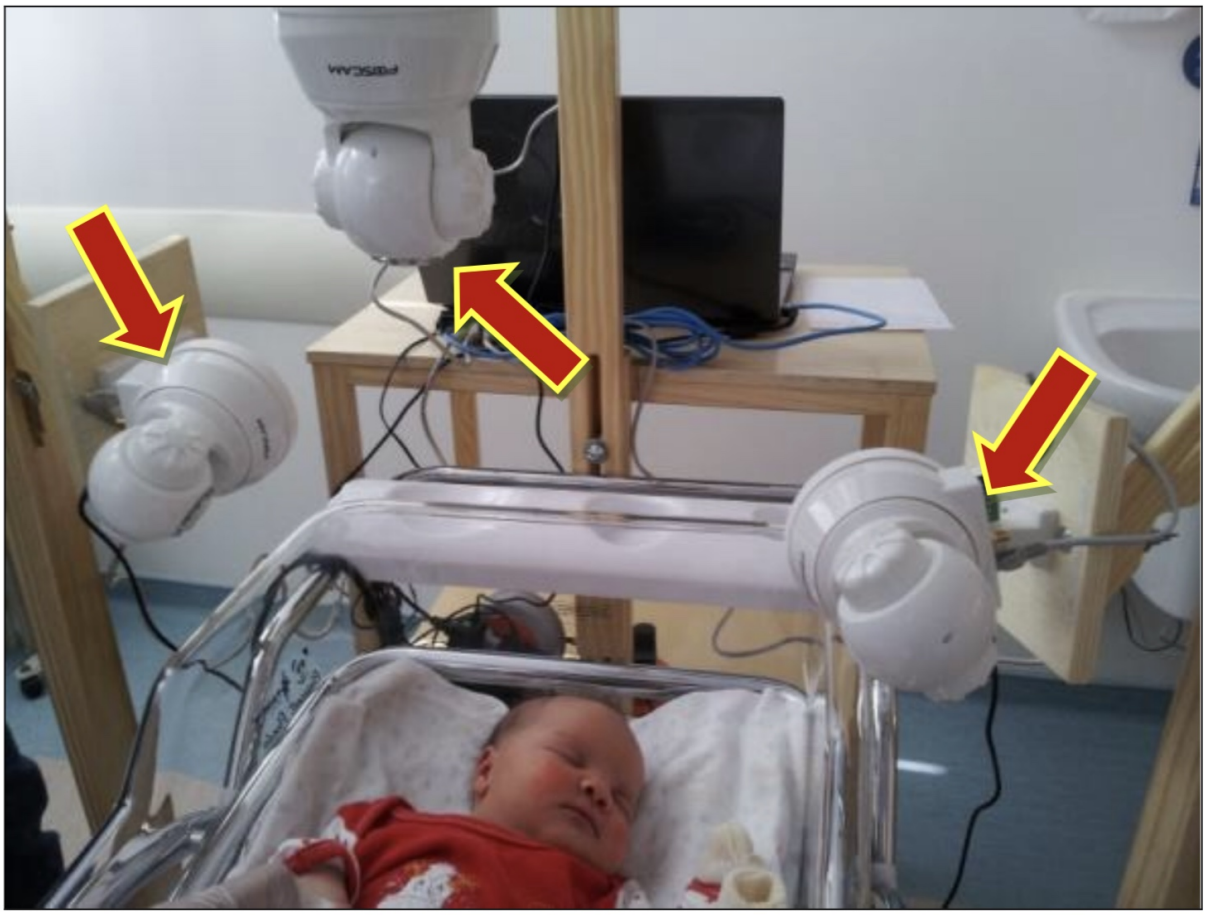
To implement the eye-tracking framework and the classification models, we used the following datasets: (1) UNIFESP Face Image Dataset (HEIDERICH; GUINSBURG; LESLIE, 2013), and (2) infant Classification of Pain Expression (iCOPE) (BRAHNAM et al., 2006, 2007).

3.1.1 UNIFESP Face Image Dataset

Heiderich, Leslie, and Guinsburg (2015) developed the UNIFESP Face Image Dataset at the Federal University of São Paulo, Brazil. It includes 30 healthy neonates (7 late preterms and 23 born at term) with 34 to 41 weeks of gestational age and 24 to 168 hours of life. For each neonate, the authors recorded 10 minutes of videos before, during, and after painful procedures. These procedures, such as venipuncture, capillary, or intramuscular injection, were performed while collecting routine tests or administering vaccines. After capturing the videos of each neonate, the authors extracted frames every three seconds.

The authors captured all photographs using three Foscam cameras with a resolution of 320x233 (Figure 2). From the images captured by the system, the authors chose 12 images for each one of the 30 participating neonates. In total, 360 images were collected, of which: 138 captured before a painful procedure, 30 during the procedure, and 192 images captured within 10 minutes after a painful procedure. Subsequently, health professionals randomly evaluated each image. These health professionals were neonatologists with clinical experience working in neonatal intensive care units. The assessment led to 164 images classified as “pain” and 196 images classified as “no pain”. Figure 3 shows examples of this dataset.

Figure 2 – Physical structure built to collect data for the UNIFESP Face Image Dataset.



Source: Adapted from Heiderich, Guinsburg, and Leslie (2013).

It is noteworthy that these images have not been cropped or pre-processed, maintaining the original information from the recording. Besides the face of the neonate, objects related to the hospitalisation and also other parts of the body, such as the neck and hands, are present in the images. Therefore, the similarity with real situations found in NICUs is preserved.

3.1.2 infant Classification of Pain Expression

The iCOPE dataset was developed by Brahnam et al. (BRAHNAM et al., 2005, 2006, 2007) during a study at the St. John Hospital (now called Mercy Hospital) with the Neonatology Department in Missouri, USA.

The authors captured a total of 200 images from 26 neonates, 13 girls and 13 boys, all Caucasians. The age group of these neonates ranges from 18 hours to 3 days of life. Although all of them were in good health, the authors reported that six male babies were circumcised the day

Figure 3 – Examples of images of neonates taken after painful procedure from the UNIFESP Face Image Dataset.

Source: UNIFESP Face Image Dataset.

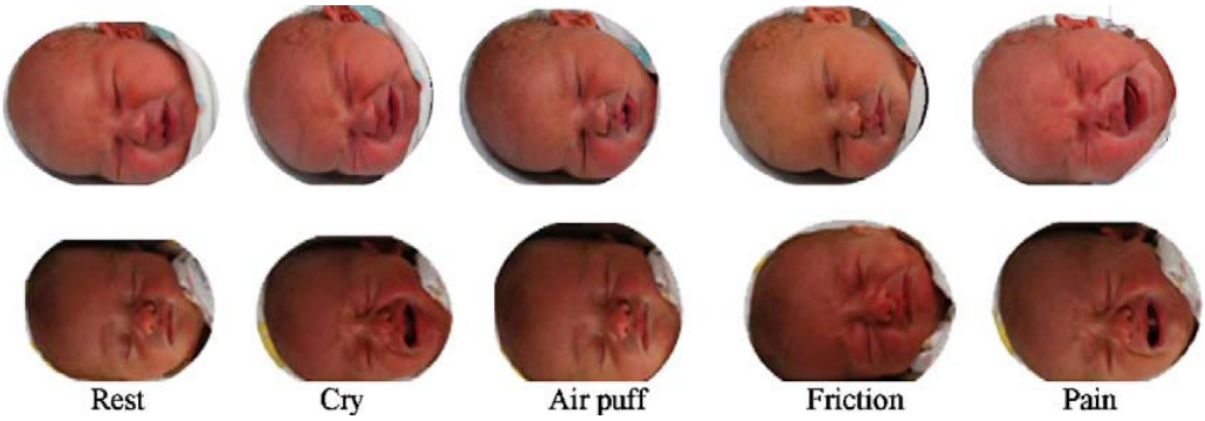
Figure 4 – Example of image of a neonate with pain from the iCOPE.

Source: iCOPE.

before the photos were captured and that everyone's last feed was done in a period of 45 minutes to 5 hours before the photographs were taken. Figure 4 illustrates an example from the iCOPE.

All images were photographed using a Nikon D100 digital camera in ambient light conditions with a resolution of 3008x2000 in a separate room from other neonates. Neonates

Figure 5 – Examples of facial expressions of the iCOPE.



Source: Adapted from Brahnam et al. (2005).

were photographed during a session in which they experienced 4 different stimuli performed in the following sequence:

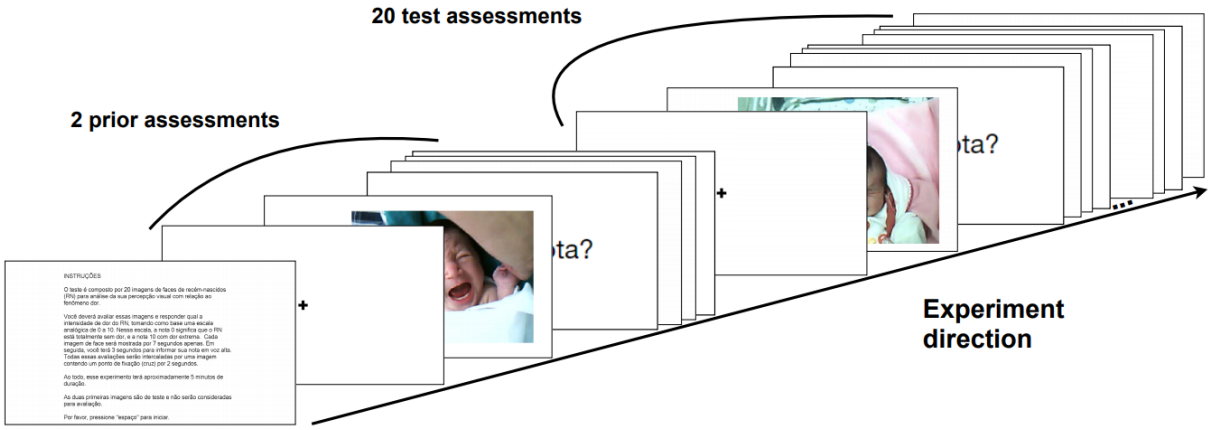
1. Transport from one crib to another: after transport between cribs, the neonate was swaddled and several photographs were taken over 1 minute. Besides, the authors observed whether the neonate was crying or resting;
2. Air stimulus: after resting for 1 minute, the neonate's nose was exposed to a breath of air emitted from a squeezable plastic camera lens cleaner;
3. Friction: after 1 minute, the external lateral surface of the heel was rubbed for 10 to 15 seconds with cotton wool soaked with 70% alcohol;
4. Pain: after 1 minute of rest, the external lateral surface of the heel was punctured for blood collection. Photographs were taken from the moment the needle was introduced until the end of the collection.

This dataset is composed of: 63 images of neonates resting, 18 crying, 23 during air stimulation, 36 during friction, and 60 during the painful procedure. Figure 5 shows a sample set of these images. However, we used only the images of neonates during rest and painful procedure to train the classification models (CARLINI et al., 2021a) .

3.2 EYE-TRACKING FRAMEWORK

To understand the human facial perception, we used previous research data of eye movements recorded with the on-screen Tobii TX300 equipment that comprises an eye tracker unit integrated to the lower part of a 23in TFT monitor with a screen resolution of 1280x1024 pixels (CARLINI et al., 2020; SOARES et al., 2021; SILVA et al., 2021; CARLINI et al.,

Figure 6 – Eye-tracking methodology.



Source: Author.

2021b; BARROS et al., 2021; TAMANAKA et al., 2022). The eye tracker performs binocular tracking at a data sampling rate of 300Hz with a minimum fixation duration of 60ms and a maximum dispersion threshold of 0.5 degrees. We considered only data from participants that the eye-tracker collected 70% or more of their gaze samples.

All data were collected at the Hospital of São Paulo, a university-affiliated hospital of the Federal University of São Paulo. Ethics Committee for Research (1299/09, 3.116.151, 3.116.146, and 3.201.307) approved this procedure, and it involved the participation of 143 volunteers divided into four groups:

a) 44 experts (physicians): 4 paediatricians and 40 neonatologists;
b) 40 health professionals: 17 nursing assistants, 10 nurses, 5 physiotherapists, and 8 speech therapists;
c) 29 newborn parents;
d) 30 laypeople: non-physicians, non-health professionals and non-parents of newborn.

Figure 6 illustrates an overview of the eye-tracking framework used (CARLINI et al., 2020). The framework is composed of the following 3 main steps: (1) an introductory screen with instructions to the participant, (2) presentation of two evaluation trials, so that the subjects learn and comprehend the experiment, and (3) the beginning of the procedure itself. Each neonatal face image to be evaluated is non-centralised located on the screen and is shown randomly to the volunteer for 7 seconds. All face images are separated in the experiment by a screen containing a cross as the central fixation point. As suggested by Holmqvist et al. (2011), this cross is shown for 2 seconds so that the ocular movement of all volunteers tends to start at the same position on the next face image screen.

Participants had to verbally answer a numerical score of the displayed image on the subsequent question screen shown for 3 seconds, using a scale ranging from 0 (no pain) to 10 (extreme pain). Overall, there were 20 frontal face images of 10 different newborns to be assessed. Each pair of images comprises one image of the newborn at rest and another image after a painful procedure, all chosen from the UNIFESP Face Image Dataset. The total time to perform the experimental procedure was approximately 5 minutes for each volunteer.

3.3 CLASSIFICATION MODELS

In this section, we describe the CNN architectures VGG-Face (PARKHI; VEDALDI; ZISSERMAN, 2015; CARLINI et al., 2021a) and N-CNN (ZAMZMI et al., 2019), and detail the training protocol that we followed to optimise both models. All computational methods were implemented using the TensorFlow framework (MARTÍN ABADI et al., 2015).

To the training of the classification models, we applied the state-of-the-art RetinaFace face detector (DENG et al., 2019) on both face image datasets. It is a single-stage pixel-wise face localisation method that employs a multitask learning strategy to simultaneously predict face score, face box, five facial landmarks, and the 3D position and correspondence of each facial pixel. RetinaFace detected all faces from UNIFESP and iCOPE datasets and generated a new cropped image focusing on the face of the neonate.

3.3.1 VGG-Face

Following the results of Zamzmi et al. (2018), we used the VGG-Face with 16 layers (PARKHI; VEDALDI; ZISSERMAN, 2015). However, differently from the Zamzmi et al. (2018) methodology, in which the deep features extracted from the last convolutional layers were used for pain classification, Carlini et al. (2021a) implemented new fully connected layers with Rectified Linear Unit (Rectified Linear Unit (ReLU)) activation for pain presence classification, while transferring the original learning of the convolutional layers.

To define experimentally the best architecture of fully connected layers in terms of accuracy, we carried out a random search using from 1 to 3 layers with 50 to 2048 neurons each (CARLINI et al., 2021a). Using both datasets, the result suggested the architecture of 2 fully connected layers with 512 neurons each, leading to an architecture with 27,823,938 total parameters. Appendix A shows the final classification model.

This random search also determined the learning rate ($\eta_{\text{prior fine-tuning}} = 1e-4$ & $\eta_{\text{during fine-tuning}} = 1e-6$) and the weight regularisation ($l_1 = 5e-4$) values. The parameters used for training are important to avoid the overfitting of the model, consequently, increasing its generalisation capability. Also, we used DropOut (50%) layers between fully connected ones. As the loss function, we used the Categorical Cross-Entropy, given by

$$Loss = -\sum_{i}^{C} y_i log(f(x)_i), \tag{1}$$

where C is the number of distinct classes, y_i is the true answer, and $f(x)_i$ is the answer given by an activation function (e.g. Softmax). Also, L_1 regulariser was applied to penalise layer parameters during optimisation. This regulariser is described as

$$\text{Penalty}_{l_1} = l_1 \sum_{i=0}^{n} |x_i|. \tag{2}$$

RMSprop was chosen as the optimiser of the training, since it presents notable results in transfer learning frameworks (HINTON; SRIVASTAVA; SWERSKY, 2012). It is noteworthy that the term that multiplies the gradient ∇w_t fluctuates. Therefore, the learning rate changes dynamically based on the training sample. The optimiser is represented by

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{rms_t} + \epsilon} \nabla w_t, \tag{3}$$

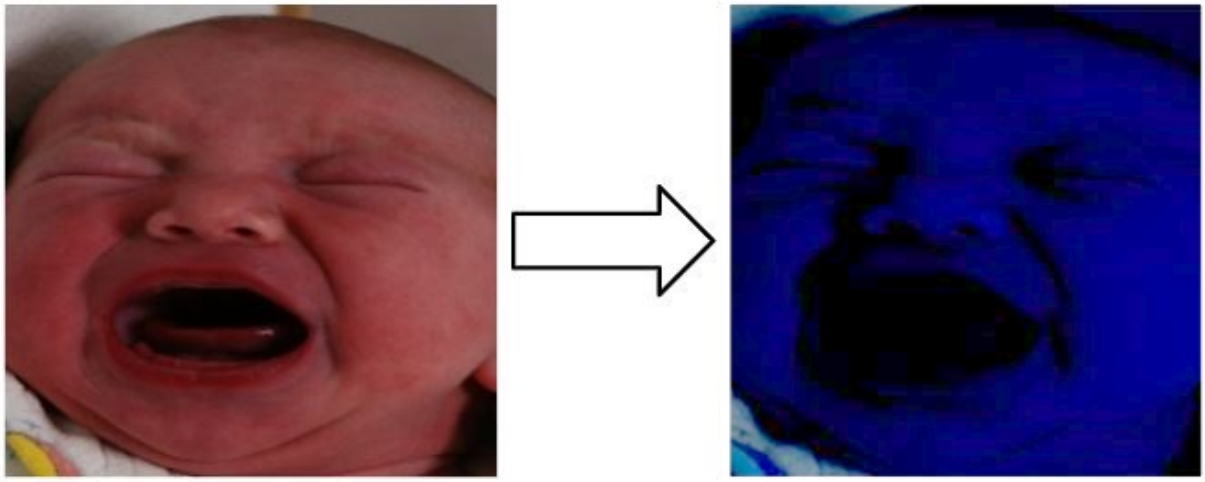
where $rms_t = \rho * rms_{t-1} + (1-\rho) * (\nabla w_t)^2$. Martín Abadi et al. (2015) suggests $\rho = 0.9$ and $\epsilon = 1e-7$.

Prior to training, we reshaped all images to the VGG-Face's input size of 224x224. Also, as required by Parkhi, Vedaldi, and Zisserman (2015), the colour channel was modified from RGB to BGR. Each channel was centralised on zero using the mean value of the original dataset used by Parkhi, Vedaldi, and Zisserman (2015). Figure 7 illustrates this step.

3.3.2 Neonatal Convolutional Neural Network (N-CNN)

Zamzmi et al. (2019) proposed the N-CNN. We adapted this original architecture, described in Section 2.2, in order to apply XAI techniques on it. As it will be highlighted in the Section 3.4, both algorithms extract feature relevance based on the classification of the model and rely on the classification index of the output. Since the original N-CNN uses a sigmoid function as the output, presenting a single classification index, we implemented, instead, a dense layer of two units with softmax activation. This adaptation resulted in a architecture with 78,234 total

Figure 7 – VGG-Face preprocessing.



Source: Author, using the iCOPE.

parameters. Appendix B shows the adapted architecture. We implemented all computational methods using the TensorFlow framework (MARTÍN ABADI et al., 2015).

It should be noted that the convolutional layers of the central and right branch are followed by Leaky ReLU activation ($\alpha = 0.01$), and DropOut (10%). After the merging layer, ReLU was applied after the convolutional and fully connected layers, and DropOut after the latter.

While the loss function and optimiser were the same as the VGG-Face's ones, the N-CNN uses L_2 regulariser with $l_2 = 0.01$, given by

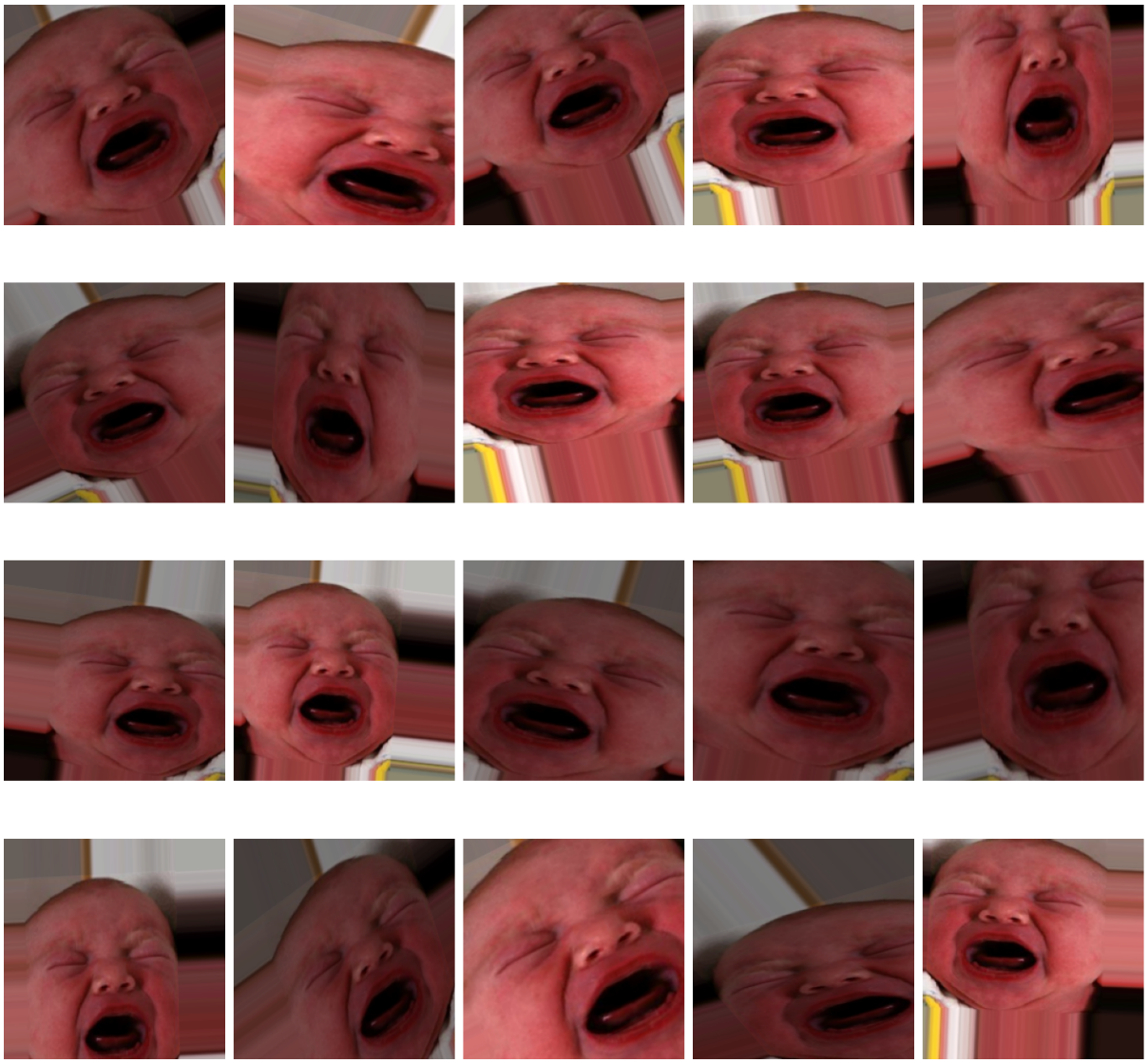
$$\text{Penalty}_{l_2} = l_2 \sum_{i=0}^{n} x_i\hat{}2. \tag{4}$$

We trained the adapted architecture end-to-end with the UNIFESP Face Image dataset and the iCOPE. Prior to training, we reshaped all images to 120x120 size, following the N-CNN's input size.

3.3.3 Training Protocol

To evaluate each model, we implement a training protocol named leave-some-subjects-out cross-validation, based on the same reasoning as the leave-one-subject-out protocol. In other words, we aggregate images based on the identity of each neonate and, then, we evaluate the model in all images regarding the neonates of the test set. Each test fold contained 10% neonates of each dataset. This way, we ensure that training and test sets are fully independent, avoiding the leaking of information and better analysing the generalisation capabilities of each model when assessing new and different neonates (COUTRIN et al., 2022). Also, we believe that this

Figure 8 – Illustration of Data Augmentation.



Source: Author, using the iCOPE.

training protocol is more suitable when transferring learning the knowledge of pre-trained CNNs. Specifically about the VGG-Face architecture, its feature extraction process is heavily based on facial recognition and may bias the actual performance of the model in pain assessment.

In order to avoid over-fitting, we applied data augmentation on both face images datasets. This procedure, widely used in the literature, aims to increase the number of available samples performing subtle changes on it (PARKHI; VEDALDI; ZISSERMAN, 2015; PEREZ; WANG, 2017; MIKOŁAJCZYK; GROCHOWSKI, 2018). To each original image from both datasets, we generated 20 new images (Figure 8) randomly using the following parameters: rotation (30º), shear (0.15), width and height shift (0.20), brightness (0.50 - 1.1), zoom (0.70 - 1.5), and horizontal flip. To verify if these images present the entire face of the neonate we double

check it with the RetinaFace algorithm. Therefore, we only consider for training the ones that RetinaFace was able to successfully detect the faces. Regarding the number of images per weight adjustment, we used a mini-batch of 16 images during training, speeding up the process and reducing over-fitting.

Regarding the VGG-Face, training started with weight adjustment of the fully connected layers, saving the weights of the model that present the lowest error. Fine-tuning started when results did not improve for 5 successive epochs, when convolutional layers of Group 4 & 5 were trained as well. The training process stopped when the error value converged. To the N-CNN's training, we did not perform fine-tuning, and the training process stopped when the error did not improve for 10 consecutive epochs.

3.4 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Artificial Intelligence algorithms, such as Deep Neural Networks, have achieved remarkable performance in a variety of tasks and contexts. However, due to their deep complex structures, it is not clear and simple how to understand the decision-making process that results in an answer. Consequently, these models are commonly known as "black boxes", where an input sample to be processed is inserted and an accurate and precise response is given at the end. Furthermore, quantitative metrics (e.g., accuracy and F1 score) are only a set of performance measures that are purely based on the actual and predicted answer to a given input. Therefore, there is a need to qualitatively understand the features that are relevant to and extracted by the decision-making process of such models. In this context, XAI techniques provide many benefits: developers can use this to debug and improve their algorithms and datasets, and end-users can understand the reasoning of a model's prediction and, ultimately, develop trust in the model (SUNDARARAJAN; TALY; YAN, 2017). Also, Selvaraju et al. (2017) state that when AI models perform significantly better than humans, the explanations provided by it should enhance human knowledge (machine teaching).

In Cheng et al. (2022), it is provided three approaches to generate explanations: directly from a transparent machine learning algorithm (e.g., Decision Trees), or from an algorithm that creates a post-hoc modelling, or visualisation tool. We benefit from the third approach, using mechanisms external to the main classification model to help describe the decision-making process of it. Techniques that apply this approach produce an attribution mask to a given input, highlighting its relevant features, such as, in our work, the pixels of the input image.

This section firstly presents in details two XAI techniques:

1. Integrated Gradients, proposed by Sundararajan, Taly, and Yan (2017);
2. Gradient-weighted Class Activation Mapping (GradCAM), proposed by Selvaraju et al. (2017);

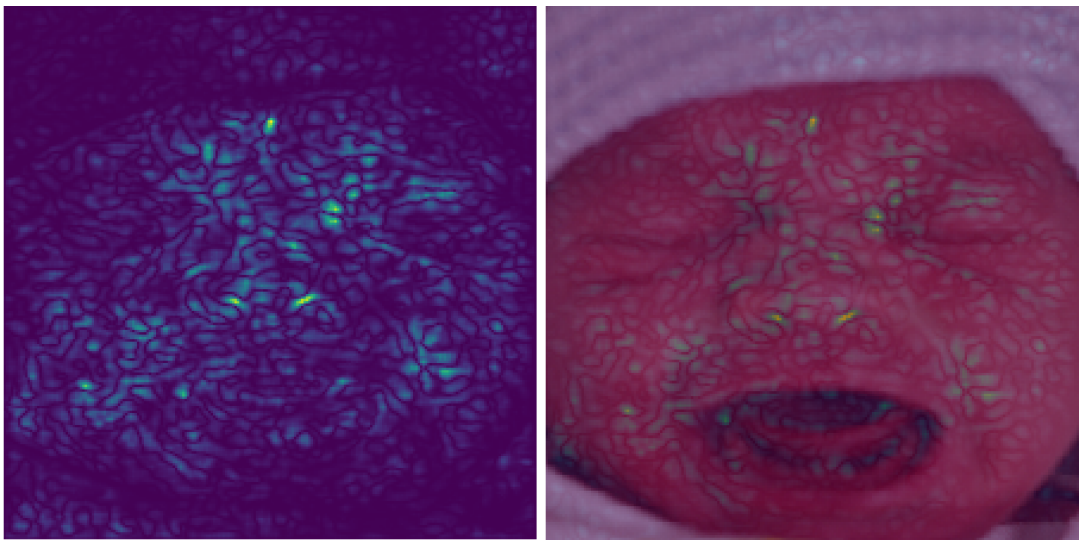
We applied both methods on all images from the UNIFESP Face Image Dataset that were used to collect eye-tracking data (CARLINI et al., 2020). TensorFlow provides a full implementation of Integrated Gradients[1], and Keras implements GradCAM[2].

Additionally, we describe in this subsection the Attribution Mask Processing, a required step to represent the attribution mask similar to human heatmaps.

3.4.1 Integrated Gradients

Figure 9 – Example of applying Integrated Gradients.

(a) Input feature interpolation in terms of γ.



(b) Attribution mask. (c) Input + attribution mask.

Source: Author, using iCOPE.

[1]Available on: `https://github.com/tensorflow/docs/blob/master/site/en/tutorials/interpretability/integrated_gradients.ipynb`.

[2]Available on: `https://keras.io/examples/vision/grad_cam/`.

Integrated Gradients is a robust XAI technique introduced by Sundararajan, Taly, and Yan (2017). The goal of the method is to accumulate the local gradients of a given pixel, assigning its importance as a score for how much it adds to the overall probability of the model's classification, and, consequently, obtaining an attribution mask.

In contrast to previous related works on XAI, Sundararajan, Taly, and Yan (2017) state that every attribution method must have some desirable characteristics (or axioms) in order to tease apart errors that stem from the misbehaviour of the model versus the misbehaviour of the attribution method. These properties are:

1. Sensitivity, this axiom has two parts:
   a) If the baseline and the input differ only in one feature, but have different predictions, then this feature gets non-zero attribution;
   b) If a feature does not play any role in the network, it receives no attributions.
2. Implementation Invariance: Two networks are functionally equivalent if their outputs are equal for all inputs, despite having very different implementations. Consequently, the attributions are always identical for two functionally equivalent networks.

The Integrated Gradients' equation combines both properties, and the attribution to a given input feature $L_{\text{IntegratedGradients}}(x_i)$ is obtained by

$$L_{\text{IntegratedGradients}}(x_i) ::= (x_i - x_i') \times \int_{\gamma=0}^{1} \frac{\partial F(x' + \gamma \times (x - x'))}{\partial x_i} d\gamma, \tag{5}$$

where x is the input, x' is the baseline, i is the feature of the input, $F(x)$ is the function (classification model), and γ is the interpolation value between x' and x. Figure 9 illustrates the method.

Also, Sundararajan, Taly, and Yan (2017) show that the Integrated Gradients presents further characteristics that benefit the method:

3. Completeness: The attributions from integrated gradients sum to the difference between the prediction scores of the input ($F(x)$) and the baseline ($F(x')$). Therefore, the prediction is entirely accounted for;
4. Linearity preservation: If a model F is a linear combination $a \times F_1 + b \times F_2$ of two networks F_1 and F_2, then a linear combination of the attributions for F_1 and F_2, with weights a and b respectively, is the attribution for the model F. In such a way, the attribution method preserves any linear logic present within a model;

5. Symmetry-Preserving: Given two features of an input x, if $x_1 = x_2$ and their corresponding baseline is $x'_1 = x'_2$, their attributions are also equal. In other words, if two features play the exact same role in the model, then they ought to receive the same attribution.

Admittedly, the method has some limitations. As pointed by Sundararajan, Taly, and Yan (2017), Integrated Gradients do not provide any insights about features interactions and combinations. And more importantly, each analysis refers only to a single input and the behaviour of the model when processing it, consequently, it does not directly offer global feature importance across an entire dataset (SUNDARARAJAN; TALY; YAN, 2017).

3.4.2 Gradient-weighted Class Activation Mapping

In 2017, Selvaraju et al. (2017) introduced the GradCAM. Their work presents a generalisation of the previous method called Class Activation Mapping (CAM) proposed by Zhou et al. (2016). CAM can solely be applied to a CNN architecture that applies global average pooling to the final convolutional feature maps, followed by a single fully connected layer that produces the predictions (softmax). To overcome this limitation, GradCAM evaluates the gradients between the final convolution layer and the desired output. In fact, GradCAM is able to produce an attribution mask in terms of any convolutional layer of the architecture. However, as stated by the authors, it is expected that the last convolutional layers has the best compromise between high-level semantics and detailed spatial information, since the neurons in these layers look for semantic class-specific information in the image.

To obtain the attribution mask L^c_{GradCAM} to a specific c class of a given input x, it is firstly needed to compute the gradient $\frac{\partial y^c}{\partial A^k_{ij}}$ of the score y_c (before the softmax) with respect to feature maps A of the desired convolutional layer k. Then, the weights importance α of each feature map are calculated by

$$\alpha^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients via backpropagation}} . \tag{6}$$

This step generalises the CAM algorithm. The original method proposed the use of the weights between the last convolutional layers and the neuron that outputs the score of a given class as the importance weights α^c_k of each feature map.

The final attribution mask is obtained by

$$L^c_{\text{GradCAM}} = \text{ReLU}\underbrace{\left(\sum_k \alpha^c_k A^k\right)}_{\text{linear combination}}. \tag{7}$$

ReLU is applied to filter out feature maps that present a negative influence on the class of study. These features are likely to influence other classes instead of the desired one.

It should be noted that, since gradients are obtained with respect to the dimensions of a specific feature map, the corresponding attribution mask L^c_{GradCAM} will present the same dimensions. A simple up-scale of the attribution mask solves this problem by matching the mask to the original input size.

3.4.3 Attribution Mask Post-Processing

In order to enhance the visualisation of each attribution mask, we propose post-processing steps on it using classical computer vision techniques. The intuition of these procedures is to enable a more direct comparison to the heatmap of the areas observed by adults during eye-tracking experiments, filtering out the less relevant pixels of the input image.

The heatmaps generated by the Tobii Studio software have a colour map with green, yellow, and red tones, meaning areas with a lesser number of fixations to more fixations (CARLINI et al., 2020). Following this, we applied this same colour map to the heatmaps that illustrate the attribution mask of both XAI methods. Figure 10a shows this first step of post-processing.

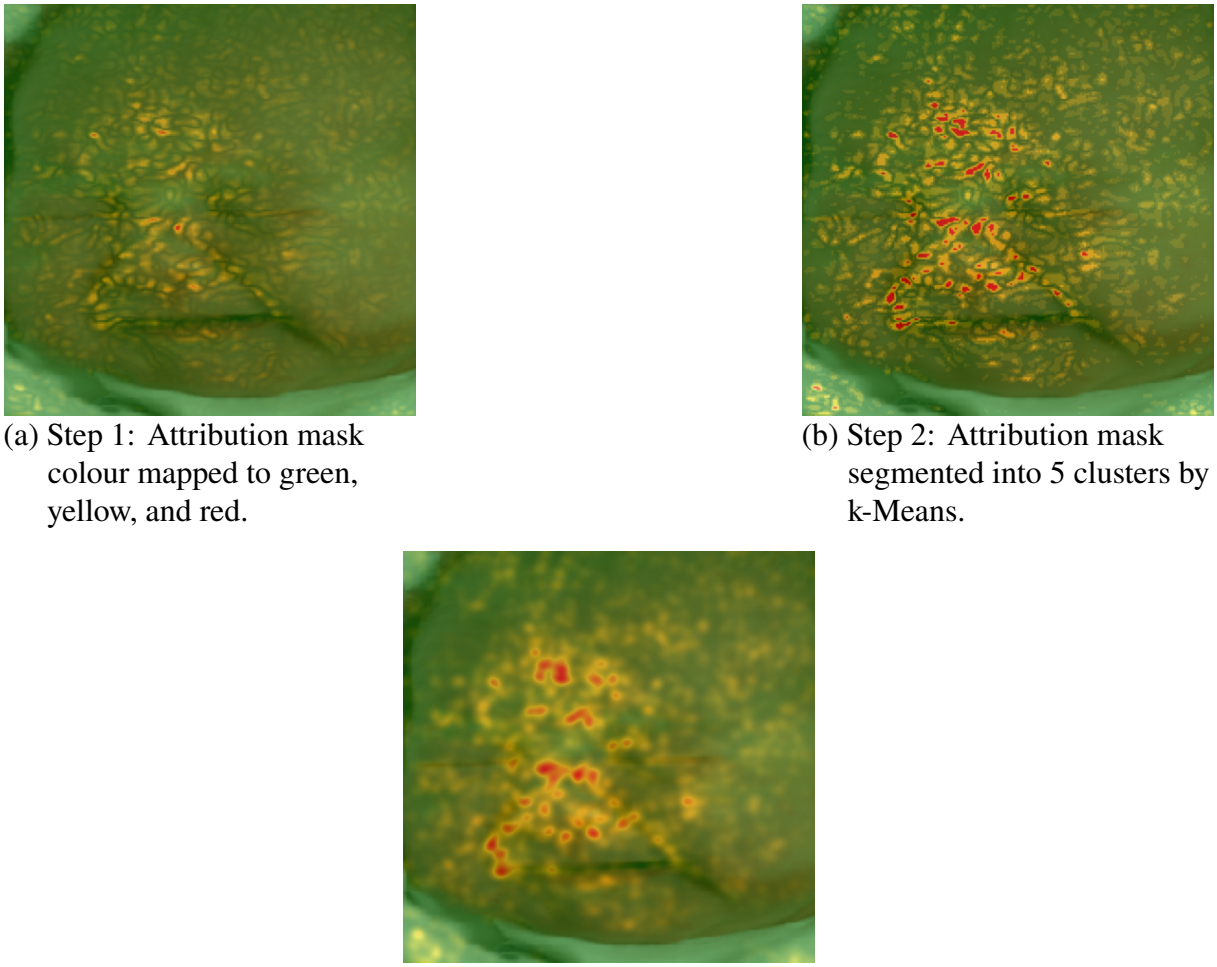
Secondly, we applied the k-Means algorithm in order to segment the attribution score into 5 clusters, enabling a better understanding of the importance of these scores. This way, clusters 1, 3 and 5 stand for green, yellow and red, and clusters 2 and 4 are the transitions between these colours. MacQueen et al. (1967) introduced the k-Means algorithm, and it aims to cluster data by trying to separate samples in n groups of equal variance, minimising a criterion known as the inertia. This can be calculated by

$$Inert. = \sum_{i=0}^{n} \min_{\mu_j \in C} \left(||x_i - \mu_j||^2\right), \tag{8}$$

where x_i is the input data, and μ_j is the corresponding mean of a given cluster τ. Finally, a new attribution mask was created by labelling each original score with its corresponding cluster mean. Figure 10b illustrates it.

In the third step, we applied a Gaussian Filter on those new attribution masks, smoothing their values, and, consequently, the transitions between colours. This is a low-pass filter that

Figure 10 – Steps 1 to 3: Integrated Gradients' attribution mask post-processing.



(a) Step 1: Attribution mask colour mapped to green, yellow, and red.

(b) Step 2: Attribution mask segmented into 5 clusters by k-Means.

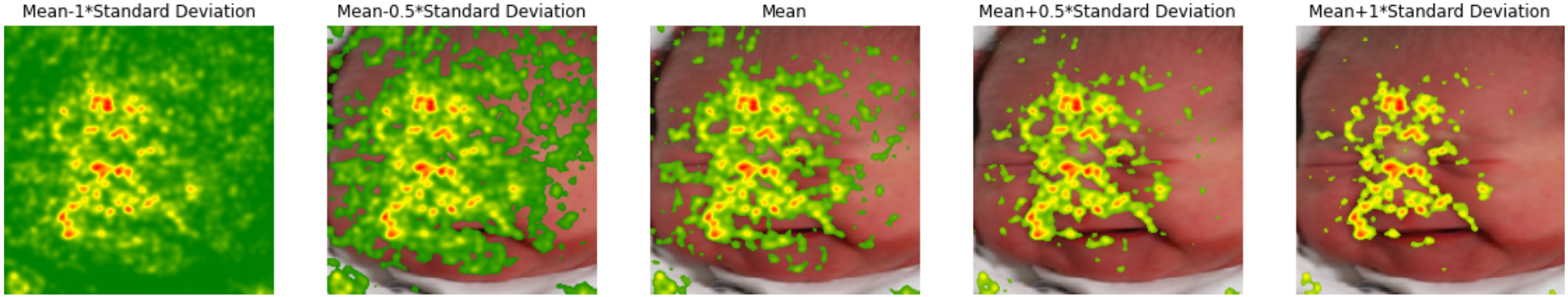
(c) Step 3: Clustered attributions smoothed by a Gaussian Filter of size 11x11.

Source: Author, using the iCOPE.

blurs an image using a Gaussian function, commonly used to describe a normal distribution. We used a kernel of dimensions 11x11 with a standard deviation equal to 2. Figure 10c shows an example of this step.

As the last step of post-processing, we computed an alpha channel to filter out less relevant attribution scores. Alpha channels are applied when dealing with images composed of many layers (e.g. an image of neonate + heatmap of the attribution mask). The main idea is to assign a numeric value to each pixel of the up-front layer, ranging from 0 to 1. A value of 0 means that the pixel is fully transparent, and the image beneath this up-front layer will be shown. A value of 1 means that the pixel is fully opaque. We assigned these values (0 or 1) based on the mean μ

Figure 11 – Step 4: Filtering out less relevant Integrated Gradients' smoothed attributions using the mean and standard deviation of it.



Source: Author, using the iCOPE.

and standard deviation σ of the attribution mask A obtained in the previous step:

$$\alpha_{channel}(A) = \begin{cases} 1 & \text{if } L_i \geq \mu_L + c \times \sigma_L \\ 0 & \text{if } L_i < \mu_L + c \times \sigma_L \end{cases}, \tag{9}$$

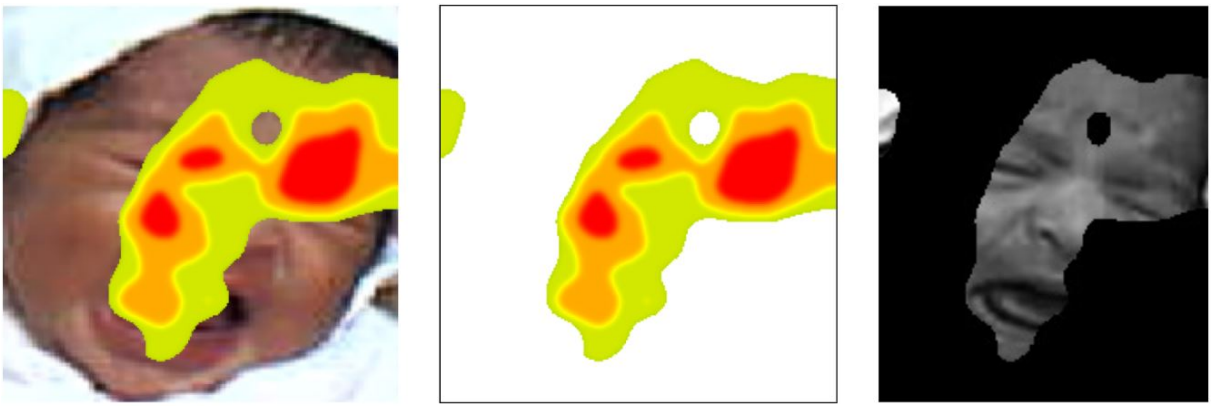
where c is a scalar, and L_i is the attribution score of a pixel of x. In this work, the heatmap of the mask is the up-front layer of the final composed image. Experimentally, we determined $c = 0.5$. Figure 11 illustrates several alpha channels applied to an attribution mask.

3.5 PERCEIVED FACIAL AREAS

Recently, Schiller et al. (2020) proposed a comparison method between heatmaps of eye-tracking data and attribution masks obtained by XAI techniques. Hwu et al. (2021) also used this framework. In the first work, the authors compared the facial perception regarding human emotions (e.g., “neutral”, “happy”, “sad”, etc) with the attribution mask of CNN models trained for automatic emotion recognition. Subsequently, Hwu et al. (2021) compared the heatmap explanations of a CNN model trained to predict driving behaviour with the gaze of human drivers.

Following Schiller et al. (2020), we created grey scale contour masks of all heatmaps generated by the adults’ facial perception and by the attribution of XAI methods. These contour masks show the most perceived regions of the image when assessing the neonatal pain.

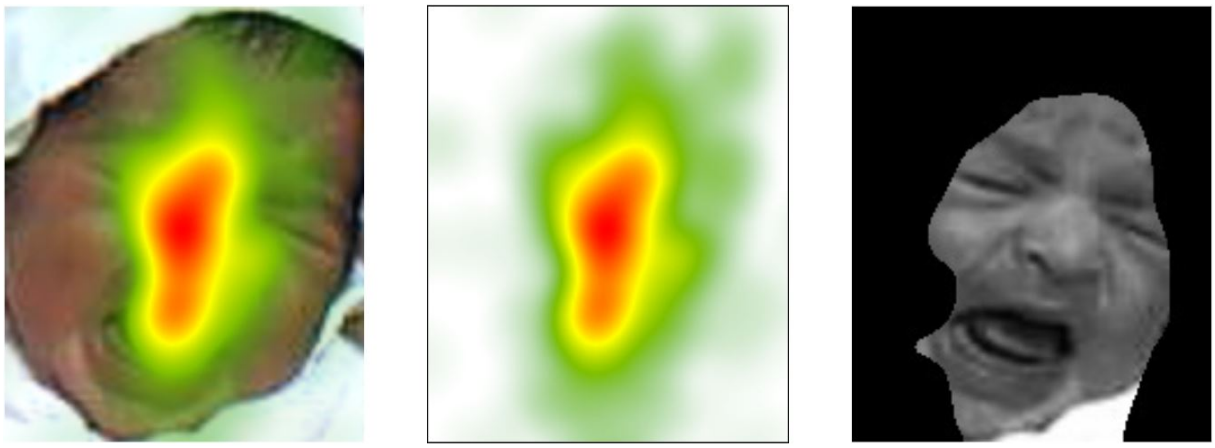
Figure 12 – Example of a contour mask of the most relevant regions according to GradCAM.



Source: Author, using the UNIFESP Face Image Dataset.

Regarding the attribution masks obtained by the XAI methods, their contour masks were created directly based on the processed attributions (Subsection 3.4.3). Figure 12 illustrates this. It should be reminded that attribution masks generated by XAI methods have the same dimensions as the input image (VGG-Face: 224x244; N-CNN: 120x120). However, in order to

Figure 13 – Example of a contour mask of the perceived areas by physicians when assessing a neonate with pain.



Source: Author, using the UNIFESP Face Image Dataset.

compare them with the adults' facial perception, the XAI's contour masks were resized to match dimensions with the adults' mask.

Regarding the adults' contour masks, we applied an alpha channel in order to filter out the less perceived areas by adults. The heatmap images generated by the Tobii Studio software were converted to grey scale and, then, filtered out using analogously Equation 9 with $c = 0.1$. Figure 13 shows an example of this procedure.

3.6 IMAGE COMPARISON METRICS

To compare the perceived facial areas and its corresponding contour masks, we used quantitative metrics widely used in the field of computer vision. These metrics were also used by Schiller et al. (2020) and Hwu et al. (2021).

The first metric is the cosine similarity, given by

$$\cos(\theta) = \frac{\mathbf{M}^{\mathbf{1}}_{\text{flat}} \cdot \mathbf{M}^{\mathbf{2}}_{\text{flat}}}{\|\mathbf{M}^{\mathbf{1}}_{\text{flat}}\| \|\mathbf{M}^{\mathbf{2}}_{\text{flat}}\|}, \tag{10}$$

where $\mathbf{M}_{\text{flat}}$ is the flattened normalised vector of the contour mask M. Since we are dealing with positive values, the cosine similarity ranges from 0 (non-similar vectors) to 1 (identical vectors).

We also used Spearman's rank correlation coefficient, described as

$$r_s = \frac{\text{cov}\left(rg_{M^1}, rg_{M^2}\right)}{\sigma_{rg_{M^1}} \sigma_{rg_{M^2}}}, \tag{11}$$

where the rg_M is the ranks of the raw values within each vector, r_s is the correlation coefficient, $\text{cov}\left(rg_{M^1}, rg_{M^2}\right)$ is the covariance of the rank variables, and $\sigma_{rg_{M^1}}$ and $\sigma_{rg_{M^2}}$ are the standard

deviations of the rank variables. Spearman's correlation is a non-parametric measure of rank correlation that ranges from -1 (negative dependency) to 1 (positive dependency). When $r_s = 0$, that means there is no dependency at all between M^1 and M^2.

The last metric used in this work is the Overlap. Proposed by Schiller et al. (2020), the result of this metric is a value between 0 and 1. A value of 1 means that both heatmaps have identical position and contour, while a value of 0 means that there is no overlap between the viewed areas at all. Mathematically, let $M^1 = (a_1,a_2,...,a_n)$ and $M^2 = (b_1,b_2,...,b_n)$ be two flattened contour masks, this metric is given by

$$Overlap(A,B) = \frac{\sum_0^n Q(a_i,b_i)}{n}, \tag{12}$$

where $Q(a_i, b_i)$ is a concordance indicator represented as

$$Q(a_i, b_i) = \begin{cases} 1 & \text{if } a_i = b_i \\ 0 & \text{if } a_i \neq b_i \end{cases} . \tag{13}$$

4 EXPERIMENTS AND RESULTS

In this chapter, we describe the experimental results carried out in this dissertation. We show the performance of the VGG-Face and the adapted N-CNN, and, then, we qualitatively investigate GradCAM's attribution masks calculated on distinct convolutional layers of the VGG-Face and N-CNN. Next, we properly present our main contributions. Firstly, the qualitative results, illustrating examples of heatmaps and contour masks of the perceived regions by adults and by our classification models. Then, tables and violin plots with the quantitative results regarding the image comparison metrics. Finally, we discuss our results following our main research questions. Specifically, we focused on the analysis of the eye-tracking data of experts (physicians) and parents when compared to the most relevant regions extracted by our classification model. Therefore, we have the following six extraction models: VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts.

4.1 EXPERIMENTS

Prior to our main experiments, we carried out supporting ones in order to implement and obtain the classification models that will be latter used in the feature extraction, and to qualitatively investigate the feature extraction process performed by GradCAM in distinct layers of the VGG-Face and N-CNN.

4.1.1 Classification Models Performance

Table 1 indicates that the VGG-Face architecture achieved the best on average performance with 86.2% accuracy, 87.7% F1 score, and 85.4% AUC. Also, the results demonstrate that the adapted VGG-Face model presents high generalisation capabilities, as shown by the 85.9% precision and 90.3% sensibility. It should be noted that the standard deviation of all metrics is considerably high ($> 5\%$). The best result was obtained by the first training fold, where it presented 96.6% accuracy, 96.9% F1 score, and 96.4% AUC. The first training fold was chosen to be used in the latter feature extraction experiments.

As shown in Table 2, the adapted N-CNN architecture achieved an on average performance of 77.1% accuracy, 80.8% F1 score, and 76.0% AUC. Analysing the average 89.0% sensibility, we can observe that the model was able to correctly classify most of the positive labels of "pain"

Table 1 – Performance of the VGG-Face architecture.

Fold	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	Average
Accuracy	96.6%	84.0%	72.7%	92.0%	92.5%	91.1%	84.3%	76.0%	83.3%	89.6%	86.2% ± 7%
Precision	93.9%	82.6%	79.3%	95.7%	94.1%	84.4%	88.0%	71.4%	82.8%	86.7%	85.9% ± 7%
Sensibility	100.0%	82.6%	71.9%	88.0%	94.1%	100.0%	81.5%	100.0%	88.9%	96.3%	90.3% ± 9%
F1 score	96.9%	82.6%	75.4%	91.7%	94.1%	91.5%	84.6%	83.3%	85.7%	91.2%	87.7% ± 6%
AUC	96.4%	83.9%	72.9%	92.0%	91.8%	91.4%	84.5%	70.0%	82.5%	88.6%	85.4% ± 8%

Table 2 – Performance of the adapted N-CNN architecture.

Fold	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	Average
Accuracy	79.7%	72.0%	81.8%	76.0%	66.0%	71.4%	90.2%	82.0%	79.2%	72.9%	77.1% ± 7%
Precision	73.2%	63.6%	82.4%	67.6%	73.5%	69.0%	92.3%	78.4%	74.3%	71.9%	74.6% ± 8%
Sensibility	96.8%	91.3%	87.5%	100.0%	73.5%	74.1%	88.9%	96.7%	96.3%	85.2%	89.0% ± 9%
F1 score	83.3%	75.0%	84.9%	80.7%	73.5%	71.4%	90.6%	86.6%	83.9%	78.0%	80.8% ± 6%
AUC	78.7%	73.4%	80.7%	76.0%	63.1%	71.5%	90.3%	78.3%	76.7%	71.2%	76.0% ± 7%

and "no pain", therefore, presenting a low number of false negatives. However, the 74.6% precision indicates that the model presented an increased number of false positives of both classes. Consequently, it reduced the generalisation capabilities of the model, as expressed by the F1 score and AUC. Again, the standard deviation of all metrics is high ($> 5\%$). The N-CNN achieved its best performance with the seventh training fold, presenting 90.2% accuracy, 90.6% F1 score, and 90.3% AUC. Therefore, this training fold was latter used in the feature extraction experiments carried out in this work. Finally, it is noteworthy that, when assessing the 20 images from the eye-tracking experiments, the N-CNN misclassified 2 images originally labelled as "no pain" (UNIFESP dataset ID: 06-12348-145026 and 20-122223).

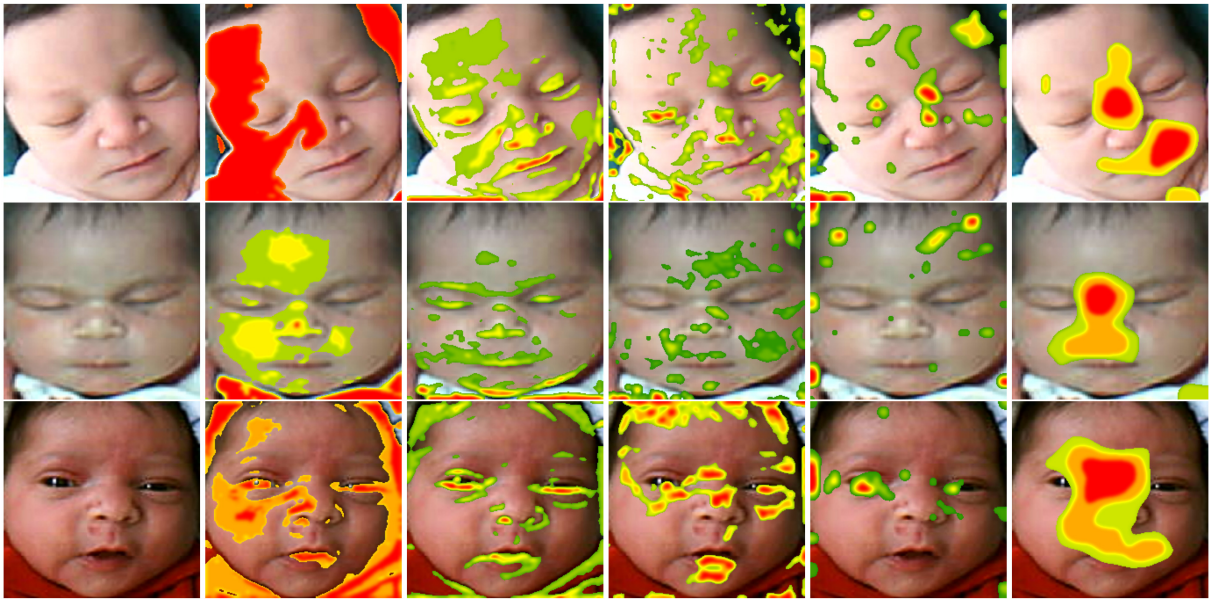
4.1.2 Layers' Depth Impact on GradCAM feature extraction

To better understand GradCAM's feature extraction described on the previous chapter, we analysed it on different convolutional layers. In other words, following Equations 6 and 7, we compute the weights α based on the gradients at different depths k of the CNN's architecture, retrieving the importance of each feature map A of the given layer.

Subfigures 14a (all images on Appendix C) and 14b (Appendix D) illustrate some examples of the attribution mask calculated on the last convolutional layer of each group of the VGG-Face architecture. The first convolutional group mainly identifies regions outside the face and some edges of it, such as the contour of the eyes and mouth. It suggests that the first convolutional group extracts information related to the contrast of the input image. The second convolutional group appears to focus more on the face than the previous layer, also highlighting contours related to contrast. The third group highlights similar regions to the second layer's one.

Figure 14 – Samples of input images and GradCAM's feature extraction on the last convolutional layer of each group of the VGG-Face.

(a) "Pain" images.



(b) "No pain" images.

From left to right: input image, 1st convolutional group, 2nd group, 3rd group, 4th group, and 5th group.

Source: Author, using the UNIFESP Face Image Dataset.

Surprisingly, the fourth group extracts granular information from several and distinct regions of "no pain" images, not having any clear pattern of extraction or being similar to the previous layers. To "pain" images, it seems that GradCAM highlighted regions extracted in the previous layers as

Figure 15 – Samples of input images and GradCAM's feature extraction on the merging and final convolutional layer of the N-CNN when assessing "no pain" (left) and "pain" (right) images.

From left to right: "no pain" input image, merge layer, final convolutional layer, "pain" input image, merge layer, and final convolutional layer.

Source: Author, using the UNIFESP Face Image Dataset.

well. Then, the fifth and final convolutional group extracts clustering information from the face. Interestingly, the regions highlighted by this group were not highlighted by the previous ones.

In Figure 15 (Appendix E), we can observe GradCAM's examples based on the merging layer and on the final convolutional layer of the adapted N-CNN. Regarding "no pain" images, the merging layer seems to mainly focus on the forehead. Analysing the final layer, it highlighted distinct facial regions. Investigating "pain" images, both layers presented similar feature extractions, highlighting the regions of the mouth. Yet, these layers also highlighted regions outside the face, such as the clothing of the neonate, and the contour of the face.

For both classification models, we chose the last convolutional layer to be investigated in our main results.

4.2 RESULTS

In this section, we present our main results. Firstly, we qualitatively compare the facial perception of all extraction models. Then, we investigate the quantitative results of the image comparison metrics.

4.2.1 Qualitative Results

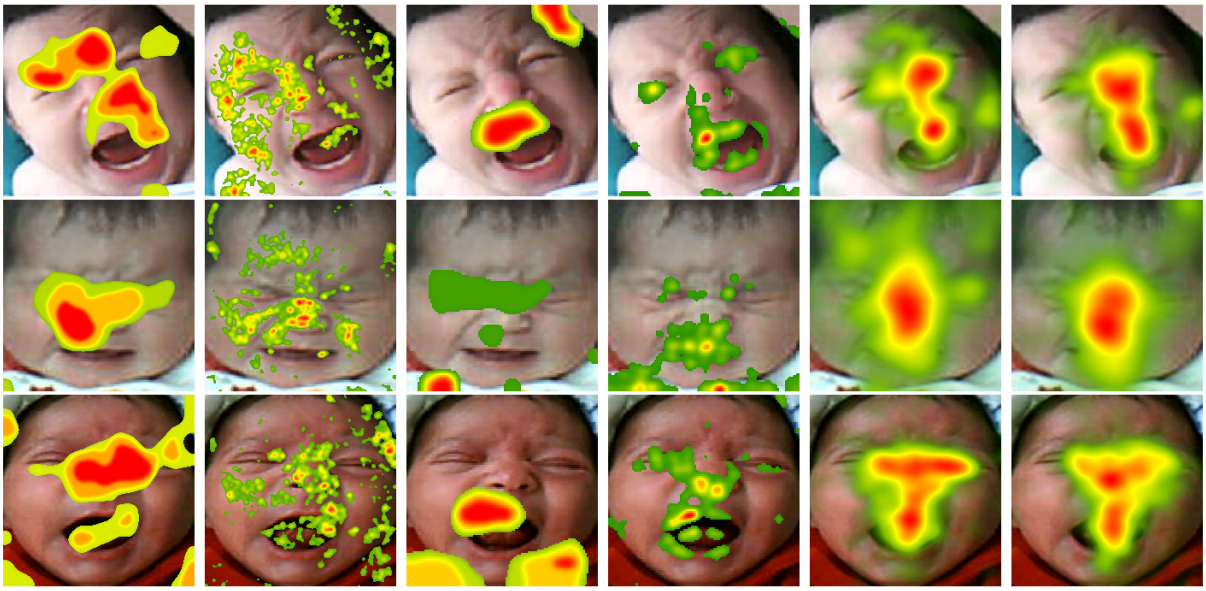
Subfigure 16a (all images on Appendix F) shows examples of the heatmaps of each extraction model when analysing "pain" images. The results of "no pain" images are shown in Subfigure 16b (Appendix H). The corresponding contour masks are shown in Appendices G and I. The columns of these figures illustrate the facial perception of the VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts, respectively.

Analysing VGG-Face's GradCAM applied on "pain" images, we can observe that it presented a clustering feature extraction in specific facials regions. These regions are mostly the forehead, eyes, nose, and mouth. Meanwhile, Integrated Gradients extracted granular information from the forehead, contour of the eyes, and the nose. In some samples, it also extracted information from the region of the mouth and nasolabial furrow, including some artefacts of the image that are not task-specific, such as the blanket and hair of the neonate, for instance. Comparing both XAI techniques, they did not present a similar feature extraction, even though both methods were applied on the same classification model. Each XAI algorithm extracted different information from the neonate's face. Furthermore, the intensity of the heatmaps by GradCAM is different of the Integrated Gradients' ones. The latter presents a higher amplitude of the attribution score, illustrated by the variation of green to red colour tones of the heatmap. GradCAM's heatmaps are majorly in red and warmer tones.

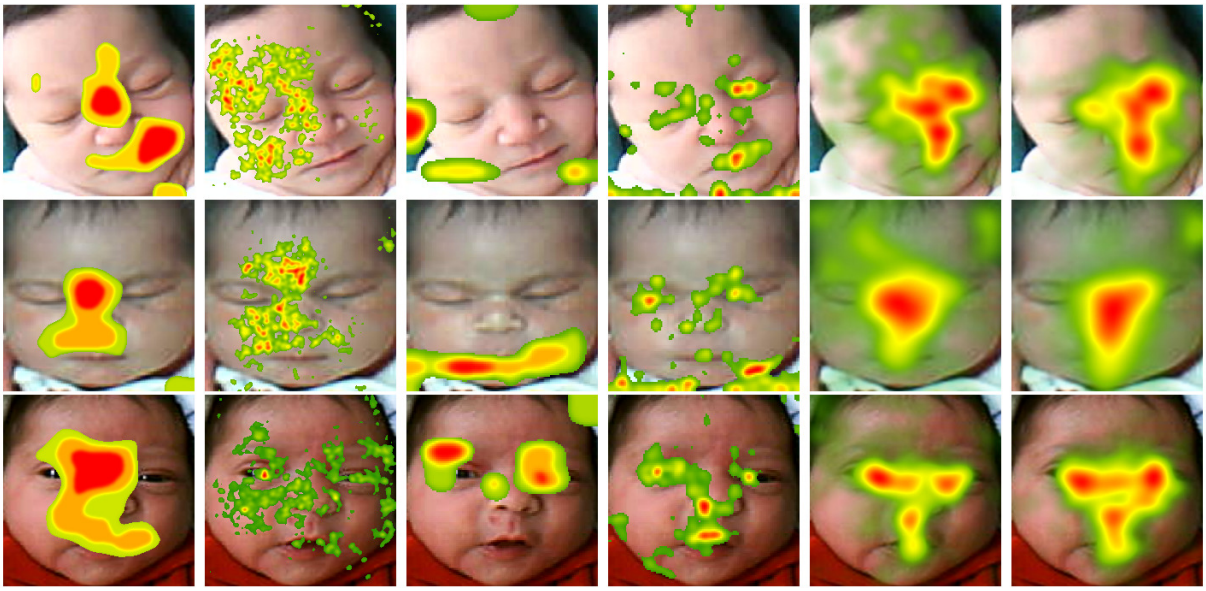
Next, analysing N-CNN's GradCAM applied on "pain" images, the model extracted clustering information from the mouth in almost all "pain" images, as shown by the red tones. Integrated Gradients also extracted granular information from this region, but it also highlighted other regions, such as the eye and forehead, with a lower attribution score. In some images, the N-CNN also extracted non-task-specific information, such as the contour of the face. Differently from what was observed in the VGG-Face's results, both XAI methods extracted similar information, mainly focusing on the mouth, even though Integrated Gradients highlighted other regions as well.

Investigating humans when perceiving "pain" images, Parents and Experts extracted holistic and similar information from the neonate's face, with no clear differences between them. However, Experts seem to present a more focused facial perception when compared to Parents. This can be visualised by the sparse and light green tones of the heatmap and the larger regions of the perceived facial areas. Comparing adults' model with machine's ones, the clustering feature

Figure 16 – Samples of heatmaps of facial features extracted by VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts.



(a) "Pain" images.



(b) "No pain" images.

Source: Author, using the UNIFESP Face Image Dataset.

extraction of the VGG-Face's GradCAM seems to present the most similar facial perception to the holistic extraction of Parents and Experts.

Regarding "no pain" images, VGG-Face's GradCAM also resulted in a clustering behaviour of perceived facial areas. However, this time, it seems that GradCAM focused mostly on centred facial regions, named the region between eyebrows, nose, and mouth. Analogously Integrated Gradients also focused on these central regions, extracting granular information from them.

Curiously, Integrated Gradients highlights the contour region of the eyes, but not the eyes itself. Both XAI methods seem to agree more to “no pain” images than to “pain” ones, highlighting similar regions. However, they presented dissimilar intensities of attribution scores. In other words, the relevance of the highlighted regions is different to each XAI technique.

When analysing N-CNN’s GradCAM assessing “no pain” images, we did not observe any clear pattern of facial feature extraction and intensity of attribution score. Also, in almost all images, it highlighted non-task-specific regions. Meanwhile, Integrated Gradients mainly highlighted the mouth, and eyes (and its eyelids). Curiously again, it also highlighted the nostril of the neonate in some images. The higher attribution scores seem to be on the eyes and mouth regions. Different from the observed information on “pain” images, the XAI methods applied on the N-CNN analysing “no pain” images appear to disagree between them, highlighting different regions, and scoring distinct attribution relevances as well.

As expected, Parents and Experts did not highlight distinct facial regions when comparing them. Their holistic facial perceptions were focused on the regions of the eyes, nose, and mouth. Again, Experts had a more concentrated facial perception than Parents. Finally, VGG-Face’s GradCAM once more seems to present the most similar facial perception to the Parents and Experts’ one, showing a clustering and holistic facial perception.

4.2.2 Quantitative Results

In this section, we show the results of the image comparison metrics between all extraction model. Since we have 6 models, there are 15 distinct pairs of comparison to be analysed. To each evaluation metric, we have:

$$10 \text{ neonates} \times 2 \text{ classes} \times 15 \text{ pairs of comparison} = 300 \text{ sample results.} \tag{14}$$

Tables 3 and 4 show the average of each evaluation metric to “pain” and “no pain” images, respectively. Table 5 exhibits the statistical difference when comparing the agreement between each extraction model towards “pain” and “no pain” images. Prior to testing statistical differences, we performed the Shapiro-Wilk normality test in all evaluation metric distributions (SHAPHIRO; WILK, 1965). Razali, Wah, et al. (2011) suggest that the Shapiro-Wilk test is the most suitable method for small sample size problems. According to the normality test results, we applied the Student’s t-test (STUDENT, 1908) to normal distributions and the Mann–Whitney U test (MANN; WHITNEY, 1947) to non-normal ones in order to verify statistical difference at $p < 0.05$ level.

Table 3 – Average result of each evaluation metric of all perceptions towards “pain” images.

Cosine Similarity	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	1.0000	**0.3670**	**0.5812**	**0.3865**	0.3930	**0.7505**
GradCAM|N-CNN	0.3670	1.0000	0.3242	0.3319	0.1893	0.3546
GradCAM|VGG-Face	0.5812	0.3242	1.0000	0.3224	0.3244	0.5618
I.G.|N-CNN	0.3865	0.3319	0.3224	1.0000	0.2511	0.3607
I.G.|VGG-Face	0.3930	0.1893	0.3244	0.2511	1.0000	0.4202
Parents	**0.7505**	0.3546	0.5618	0.3607	**0.4202**	1.0000
Spearman	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	1.0000	0.1953	**0.4272**	**0.2584**	0.1925	**0.6161**
GradCAM|N-CNN	0.1953	1.0000	0.2285	0.2494	0.0518	0.1545
GradCAM|VGG-Face	0.4272	0.2285	1.0000	0.2281	0.1643	0.3959
I.G.|N-CNN	0.2584	**0.2494**	0.2281	1.0000	0.0635	0.2073
I.G.|VGG-Face	0.1925	0.0518	0.1643	0.0635	1.0000	0.2048
Parents	**0.6161**	0.1545	0.3959	0.2073	**0.2048**	1.0000
Overlap	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	1.0000	0.6376	0.7336	0.6651	0.6115	**0.8167**
GradCAM|N-CNN	0.6376	1.0000	**0.7581**	**0.8016**	**0.7229**	0.5979
GradCAM|VGG-Face	0.7336	0.7581	1.0000	0.7484	0.7049	0.6974
I.G.|N-CNN	0.6651	**0.8016**	0.7484	1.0000	0.7099	0.6230
I.G.|VGG-Face	0.6115	0.7229	0.7049	0.7099	1.0000	0.5894
Parents	**0.8167**	0.5979	0.6974	0.6230	0.5894	1.0000

Obs: The best correlation of each model and metric are in bold.

Table 4 – Average result of each evaluation metric of all perceptions towards “no pain” images.

Cosine Similarity	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	1.0000	**0.2230**	**0.5784**	0.4055	0.4908	**0.7600**
GradCAM|N-CNN	0.2230	1.0000	0.1457	0.1892	0.1633	0.2055
GradCAM|VGG-Face	0.5784	0.1457	1.0000	**0.4222**	**0.5022**	0.5490
I.G.|N-CNN	0.4055	0.1892	0.4222	1.0000	0.2704	0.3905
I.G.|VGG-Face	0.4908	0.1633	0.5022	0.2704	1.0000	0.4831
Parents	**0.7600**	0.2055	0.5490	0.3905	0.4831	1.0000
Spearman	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	1.0000	-0.0664	**0.4346**	0.2110	0.3303	**0.5888**
GradCAM|N-CNN	-0.0664	1.0000	-0.0873	-0.0249	-0.0595	-0.1327
GradCAM|VGG-Face	0.4346	-0.0873	1.0000	**0.2606**	**0.3393**	0.3824
I.G.|N-CNN	0.2110	**-0.0249**	0.2606	1.0000	0.0743	0.1695
I.G.|VGG-Face	0.3303	-0.0595	0.3393	0.0743	1.0000	0.3146
Parents	**0.5888**	-0.1327	0.3824	0.1695	0.3146	1.0000
Overlap	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	1.0000	0.5137	0.7033	0.6355	0.6317	**0.7714**
GradCAM|N-CNN	0.5137	1.0000	0.6364	0.6789	0.6556	0.4419
GradCAM|VGG-Face	0.7033	0.6364	1.0000	**0.7529**	**0.7605**	0.6452
I.G.|N-CNN	0.6355	**0.6789**	0.7529	1.0000	0.6990	0.5800
I.G.|VGG-Face	0.6317	0.6556	**0.7605**	0.6990	1.0000	0.5806
Parents	**0.7714**	0.4419	0.6452	0.5800	0.5806	1.0000

Obs: The best correlation of each model and metric are in bold.

Table 5 – p-Value of the statistical difference between "pain" and "no pain" distributions of the comparison of all perceptions.

Cosine	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	-	**0.0116**	0.9335	0.5815	0.1353	0.7337
GradCAM|N-CNN	**0.0116**	-	**0.0098**	**0.0036**	0.8501	**0.0046**
GradCAM|VGG-Face	0.9335	**0.0098**	-	**0.0098**	**0.0002**	0.3847
I.G.|N-CNN	0.5815	**0.0036**	**0.0098**	-	0.6776	0.4489
I.G.|VGG-Face	0.1353	0.8501	**0.0002**	0.6776	-	0.2474
Parents	0.7337	**0.0046**	0.3847	0.4489	0.2474	-
Spearman	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	-	**0.0025**	0.8719	0.2142	0.1217	0.5057
GradCAM|N-CNN	**0.0025**	-	**0.0008**	**0.0000**	**0.0450**	**0.0022**
GradCAM|VGG-Face	0.8719	**0.0008**	-	0.4528	**0.0025**	0.3847
I.G.|N-CNN	0.2142	**0.0000**	0.4528	-	0.7344	0.4423
I.G.|VGG-Face	0.1217	**0.0450**	**0.0025**	0.7344	-	0.1911
Parents	0.5057	**0.0022**	0.3847	0.4423	0.1911	-
Overlap	Experts	GradCAM|N-CNN	GradCAM|VGG-Face	I.G.|N-CNN	I.G.|VGG-Face	Parents
Experts	-	**0.0026**	0.1458	0.1270	0.4722	0.0579
GradCAM|N-CNN	**0.0026**	--	**0.0016**	**0.0001**	**0.0046**	**0.0002**
GradCAM|VGG-Face	0.1458	**0.0016**	-	0.8072	**0.0176**	**0.0312**
I.G.|N-CNN	0.1270	**0.0001**	0.8072	-	0.3901	0.1720
I.G.|VGG-Face	0.4722	**0.0046**	**0.0176**	0.3901	-	0.7491
Parents	0.0579	**0.0002**	**0.0312**	0.1720	0.7491	-

Obs: $p < 0.05$ results are in bold.

More specifically, Table 3 shows the results regarding "pain" images. Firstly analysing Experts, the highest agreement was achieved with Parents, where they presented a strong cosine ($avg_{cosine} \geq 0.70$), Spearman ($avg_{Spearman} \geq 0.60$), and overlap ($avg_{overlap} \geq 0.70$) correlations (0.7505, 0.6161, and 0.8167, respectively). The results regarding Parents were analogously, where they achieved the highest agreement with Experts. Then, N-CNN's GradCAM obtained their top cosine and Spearman scores with Experts (0.3670) and N-CNN's Integrated Gradients (0.2494), respectively, yet these correlations were weak ($avg_{cosine} \leq 0.50$, and $avg_{Spearman} \leq 0.40$). Meanwhile, N-CNN's GradCAM achieved its strongest overlap correlation with N-CNN's Integrated Gradients (0.8016). Analysing VGG-Face's GradCAM, the model acquired the highest cosine and Spearman correlation with Experts (0.5812, and 0.4272), showing moderate correlations ($0.50 < avg_{cosine} < 0.70$, and $0.40 < avg_{Spearman} < 0.60$). The top overlap score was achieved with N-CNN's GradCAM (0.7581). Subsequently, N-CNN's Integrated Gradients acquired it highest, yet weak, cosine and Spearman scores with Experts (0.3865, and 0.2584, respectively). The strongest overlap was obtained with N-CNN's GradCAM (0.8016). Lastly, VGG-Face's Integrated Gradients scored the highest cosine and Spearman correlation with Parents (0.4202, and 0.2048), even so, these correlations were weak. To the overlap, VGG-Face's Integrated Gradients acquired the best result with N-CNN's GradCAM (0.7229).

Table 4 shows the results regarding "no pain" images. Equally, Experts and Parents agreed more between themselves than with the remaining models. They presented strong cosine, moderate Spearman, and strong overlap correlations (0.7600, 0.5888, and 0.7714, respectively). Examining N-CNN's GradCAM, its top scores were the weak cosine with Experts (0.2230), and the weak Spearman and moderate ($0.50 < avg_{overlap} < 0.70$) overlap correlations with N-CNN's Integrated Gradients (-0.0249, and 0.6789). Then, VGG-Face's GradCAM achieved its best cosine and Spearman score when compared to Experts (0.5784, and 0.4346), showing moderate correlation. To the overlap metric, the model obtained the strongest correlation with VGG-Face's Integrated Gradients (0.7605). Posteriorly, N-CNN's Integrated Gradients obtained the highest agreement with VGG-Face's GradCAM (0.4222, 0.2606, and 0.7529), demonstrating weak cosine, weak Spearman, and strong overlap correlations. Similarly, VGG-Face's Integrated Gradients also acquired its best scores with VGG-Face's GradCAM (0.5022, 0.3393, and 0.7605), but with moderate cosine, weak Spearman, and strong overlap correlations.

Specifically investigating the differences of agreement between "pain" and "no pain" images, Table 5 presents the p-value of this statistical test. Firstly investigating Experts, they presented statistical difference with N-CNN's GradCAM in all metrics ("pain")[1]. Then, N-CNN's GradCAM had statistical difference with all models and metrics ("pain"), except for the cosine with VGG-Face's Integrated Gradients. Meanwhile, VGG-Face's GradCAM showed statistical difference in the cosine similarity with N-CNN's GradCAM ("pain"), N-CNN's Integrated Gradients ("no pain"), and VGG-Face's Integrated Gradients ("no pain"). To the Spearman correlation, VGG-Face's GradCAM presented difference with N-CNN's GradCAM ("pain"), and VGG-Face's Integrated Gradients ("no pain"). About the overlap, it obtained statistical difference with N-CNN's GradCAM ("pain"), VGG-Face's Integrated Gradients ("no pain"), and Parents ("pain"). Regarding N-CNN's Integrated Gradients, it presented statistical difference in the cosine similarity with N-CNN's GradCAM ("pain"), and VGG-Face's GradCAM ("no pain"). When analysing Spearman and overlap results, we found only difference with N-CNN's GradCAM (both metrics, "pain"). Posteriorly, VGG-Face's Integrated Gradients demonstrated statistical difference in cosine similarity only with VGG-Face's GradCAM ("no pain"). To Spearman and overlap correlations, there were differences with N-CNN's GradCAM (both metrics, "pain"), and VGG-Face's GradCAM (both metrics, "no pain"). Finally, Parents had statistical difference with N-CNN's GradCAM in all metrics ("pain"). Specifically to the overlap, it also presented differences with VGG-Face's GradCAM ("pain").

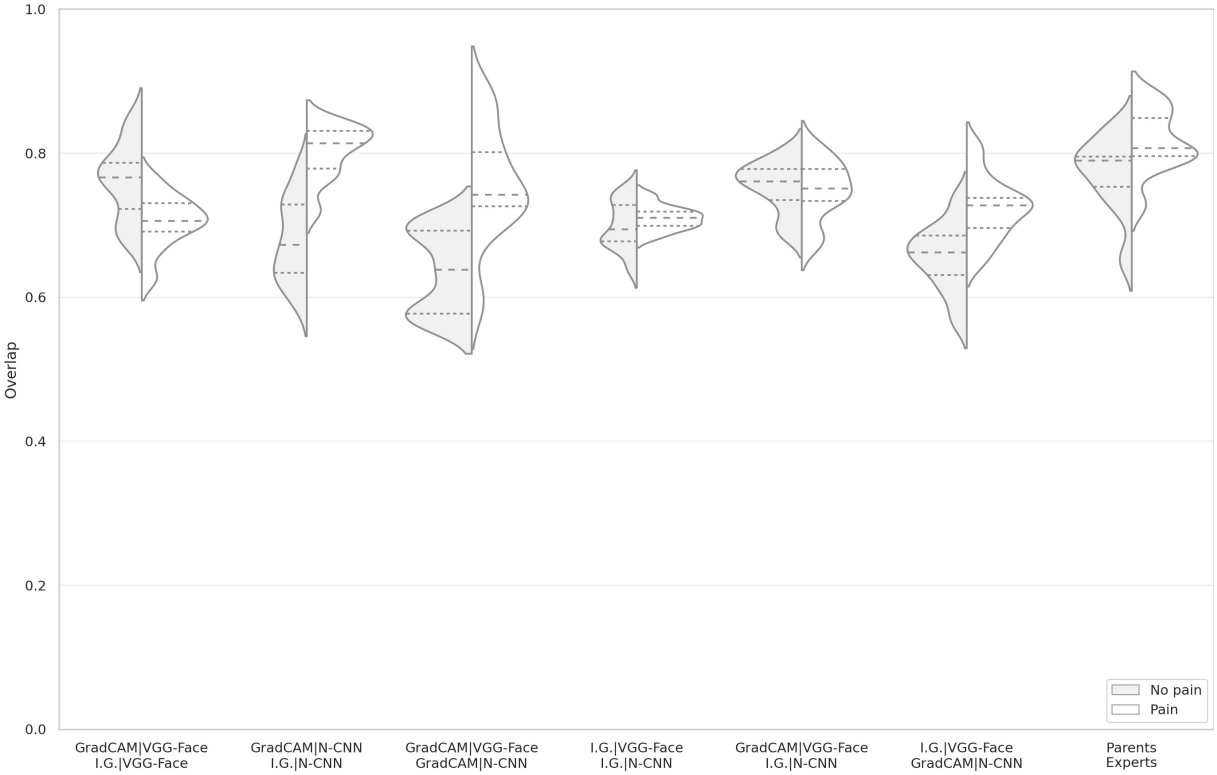
[1]For readability, the label in parentheses indicates to which class of image the pair agreed the most.

Figures 17a and 17b depict violin plots detailing the distribution of the overlap metric between all extraction models. The violin plot is similar to the popular box plot, exhibiting the median (dashed line) and interquartile ranges (dotted lines), but it also adds a probability density of the illustrated data, providing deeper insights to the analysis. For each comparison illustrated in the plot, we have a breakdown between "pain" and "no pain" images, distinguishing its probability density.

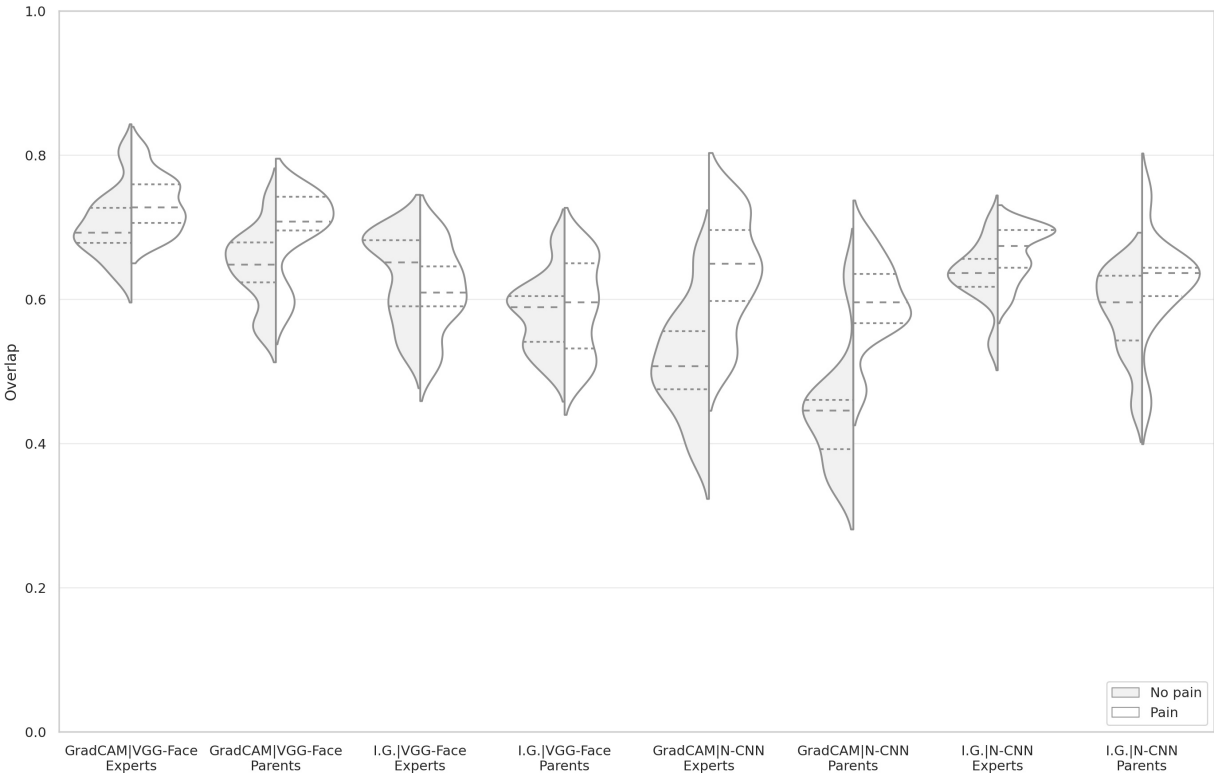
Observing Figure 17a, we can investigate the aforementioned distinctions of agreement to each class of image. VGG-Face's GradCAM and Integrated Gradients agreed more to "no pain" images, with its values mainly just below 0.8 level. More evidently, N-CNN's GradCAM and Integrated Gradients showed a higher agreement to "pain" images with average above 0.8 level, whereas the distribution of "no pain" images is spread between 0.6 and 0.8 thresholds. VGG-Face's GradCAM and N-CNN's GradCAM had a higher overlap to "pain" images, but it is noteworthy that the distribution regarding "no pain" images clearly presented two density of values: the first at 0.7 level, and the second above 0.6 level. Meanwhile, both distributions of VGG-Face's Integrated Gradients and N-CNN's Integrated Gradients were consistently at 0.7 level, but the "no pain" distribution presented slightly more spread. Then, VGG-Face's GradCAM and N-CNN's Integrated Gradients resulted in both distributions majorly between 0.7 and 0.8, with no visual distinctions between them. To the VGG-Face's Integrated Gradients and N-CNN's GradCAM, we can clearly observe that the "pain" distribution presented higher value than the "no pain" one. As expected, Parents and Experts showed an overall higher agreement than the other pairs of comparison, with distributions are at 0.8 level. Also, humans had a density of values indicating stronger overlap to a subset of "pain" images. Finally, it is also noteworthy that almost all values of the distributions shown in this figure are above 0.6 level, demonstrating moderate or strong correlations between each pair.

As shown in Figure 17b, VGG-Face's GradCAM and Experts presented moderate or strong correlation, consistently at or above 0.7 level. Also, the distribution of "pain" images has slightly greater values. Then, VGG-Face's GradCAM and Parents resulted in a "no pain" distribution majorly between 0.6 and 0.7, whereas the "pain" distribution is at 0.7 level, indicating strong correlation. Next, when analysing VGG-Face's Integrated Gradients and Experts, their "no pain" distribution clearly has higher values than the "pain" one, where the values of the latter are at 0.6 level. Regarding VGG-Face's Integrated Gradients and Parents, the "no pain" distribution is majorly concentrated at 0.6 level with moderate correlation, whereas the "pain" one is evenly spread between 0.5 and 0.7, indicating greater agreement to a subset of images. Subsequently,

Figure 17 – Violin plots showing the distributions of the overlap metric to each pair of comparison.



(a) Between each XAI method, and between Adults.



(b) Between XAI methods and Adults.

Source: Author

N-CNN's GradCAM and Experts clearly achieved a greater overlap correlation to "pain" images, with values majorly between 0.6 and 0.7. Meanwhile, the "no pain" distribution at 0.5 level also indicates slightly moderate correlation. To the N-CNN's GradCAM distributions with Parents, we observed the weak correlation to "no pain" images, where its values are below 0.5 level. Their "pain" distribution is at 0.6 level, demonstrating moderate correlation. Posteriorly, analysing N-CNN's Integrated Gradients and Experts, we observed that both distributions are between 0.6 and 0.7, indicating moderate correlation. Finally, N-CNN's Integrated Gradients and Parents presented both distributions at 0.6 level, but the "pain" achieved slightly greater values.

4.3 DISCUSSION

Overall, our findings indicate that XAI techniques present a suitable method of understanding the decision-making process performed by AI models when assessing pain through the neonatal facial expression. Also, these methods enabled the comparison of the facial perception of humans and machine towards neonatal pain assessment. Our results show that both classification models (VGG-Face and N-CNN) majorly extract information from the facial features of the neonate, but each model and XAI method presented it own decision-making process. Regarding the facial perception of Experts and Parents, the quantitative results obtained here confirm our previous findings that the human facial perception towards neonatal pain is holistic, regardless of clinical expertise and pain presence on the neonate (CARLINI et al., 2020).

Basically, we implemented and trained two CNN architectures using our proposed leave-some-subjects-out training protocol. To the VGG-Face, we applied transfer learning of the convolutional layers, and trained from scratch fully connected ones. The model achieved state-of-the-art performance, and demonstrated strong generalisation capabilities. To the N-CNN, we trained from scratch all layers. Most likely due to the lack of training samples, the model did not achieve the same performance as originally published in Zamzmi et al. (2019). Even so, all metrics are at or above 75% level. Additionally, it is noteworthy that both classification models presented a considerably high standard deviation in their average performance. We hypothesise that our training protocol led to this result, since each test fold has neonates that only appear in that fold. Consequently, there may be neonates that present a more expressive facial response to pain, or its set of image have a higher quality in terms of facial position, brightness, and resolution. Nevertheless, we believe that these conditions are necessary to properly simulate real and less controlled clinical situations.

Additionally, we investigated the GradCAM feature extraction in terms of the depth of the convolutional layer. To the VGG-Face model, GradCAM extracted non-task-related information from the image at the first three convolutional groups, such as the contour of the face or regions outside of it. And, at the fourth group, GradCAM did not present any clear pattern of extraction. However, the fifth (and last) group extracted clustering information from specific regions of the face. Meanwhile, to the N-CNN model, the merging layer and the last convolutional layer extracted similar information, but the latter seem to be more task-specific. In other words, both CNN architectures presented semantic class-specific feature extraction at the last convolutional layer. Therefore, we believe that these results agree with the assumptions presented at Selvaraju et al. (2017), where the authors suggested that this final layer presents the best compromise between high-level semantics and detailed spatial information.

Moreover, as our main contribution, we investigated the facial perception of humans and machine, and their differences, when performing neonatal pain assessment. To the VGG-Face model, the relevance of each facial feature highly depends on the applied XAI technique. In general, we observed that the VGG-Face mainly focus on the forehead, eyes, nose, and mouth when analysing "pain" images, whereas it focused on more centred regions, such as the region between eyebrows, nose, and mouth, when assessing "no pain" ones. Regarding the N-CNN model, both XAI methods suggest that the mouth is the most relevant facial region when assessing "pain" images. On the other hand, to "no pain" images, N-CNN's GradCAM did not present any clear behaviour of feature extraction. But when analysing the results provided by the merging layer, it seems to focus on the forehead. Specifically about N-CNN's Integrated Gradients, it also highlighted the eye and the nostrils of the neonate in both classes of image. Comparing both classification models, N-CNN feature extraction appears to be more focused on specific facial regions, whereas the VGG-Face extracted information from larger areas of the face. We see two possible reasons for this difference: (i) since the VGG-Face was originally trained for facial recognition using millions of images, the convolutional layers extract subject discriminant information; (ii) the N-CNN qualitative performance may be also impacted by the limited number of training samples. Additionally, the experiments carried out here did not explicitly correlate the extracted facial regions to the confidence score outputted by the model to a given class, even though both XAI methods generate their attribution mask based on the output. Consequently, a specific investigation to each facial feature is still necessary to address this question.

Regarding the comparison between GradCAM and Integrated Gradients, there is no strong agreement between them when analysing the weak values of cosine similarity and Spearman

correlation, regardless of the classification model. On the other hand, the average overlap between methods was considerable higher than the other metrics. Applied on the VGG-Face, the XAI methods presented a stronger overlap to “no pain” images. The opposite was observed regarding the N-CNN. Our hypothesis is that cosine similarity and Spearman correlation have an increased penalisation to distinct perceived regions. More importantly, it should be noted that Integrated Gradients and GradCAM have different equations to calculate the attribution mask to a given input image. Integrated Gradients individually evaluates each pixel of the input image, verifying how much it adds to the overall probability of the classification. GradCAM’s equation is based on the features maps of the last convolutional layer to a given input, weighting and averaging them to obtain the final attribution mask. Due to the processing of each convolutional layer, a unique pixel of the feature maps carries information from the corresponding neighbouring pixels of the input image as well. Also, the GradCAM’s attribution mask has the same size as the feature map, requiring an up-scale step to match with the original input size, and attributing the same score to a specific region of the image. The equations and characteristics of each method seem consistent with the attributions masks those machine models yield. Integrated Gradients extracts pixel-specific information of the input image, leading to a more granular contour mask of the perceived areas. Meanwhile, GradCAM extracts holistic and clustering information, resulting in a larger and smoother contour mask of the perceived areas. Altogether, these results raise concerns about the effective use and interpretation of XAI methods, and, more importantly, what regions of the image are truly relevant to the decision-making process of the AI model. Since several deep learning models are recently being proposed and implemented, we believe that these results stand as a cautionary warning for computer scientists and the scientific community in general for the safe and reliable use in real clinical situations.

Analysing the agreement between humans and machine models, VGG-Face’s GradCAM presented the highest comparison metrics averages within all pairs, achieving moderate and strong correlation with humans. Qualitatively, its clustering feature extraction is the most similar to the holistic human facial perception. Regarding the remaining pairs of model and XAI technique, the results suggest that the facial regions most relevant to them differs from the most observed ones by adults when assessing neonatal pain, since both classification models achieved accuracy scores comparable to adults’ ones but presented weak correlation. However, it is noteworthy the clear relevance of the mouth to the N-CNN model. As observed by Barros et al. (2021) and Silva et al. (2021), specific facial regions, such as the mouth and nasolabial furrow, are associated with a better chance of adults to correctly identify the presence of pain. Since the compared humans’

heatmaps are the average facial perception of all Experts and Parents that volunteered to the eye-tracking experiments (CARLINI et al., 2020), we believe that a region-by-region semantic comparison with volunteers that correctly identified pain presence/absence is necessary to further understand these differences and which regions lead to this result.

Lastly, we observed that the agreement between humans and machine is correlated to the label of the analysed images. According to all comparison metrics, Experts and Parents agreed more with N-CNN's GradCAM when analysing "pain" images. Specifically analysing the overlap, Parents and VGG-Face's GradCAM also presented a higher agreement to "pain" images. These results suggest that humans and machine extract distinct information when analysing "no pain" images. In other words, the discriminant facial features that define these images are different to these groups. Once again, a region-by-region comparison with adults is required.

5 CONCLUSION

This work presented novel results on the human and machine facial perception when assessing neonatal pain by disclosing a better understanding of the decision-making process performed by AI models and comparing it with the facial perception of experts and parents of neonates. In this context, our findings advance the current knowledge on how humans and machine perceive and decode the neonatal facial response to pain. We believe that our contributions enable further improvement of clinical scales and tools widely used in real situations, whether based on human or automatic frameworks.

To discriminate the presence of pain, both machine models extracted information from facial regions clinically relevant, such as forehead, eyes, nose, and mouth. Interestingly, the N-CNN architecture attributed a clear relevance to the region of the mouth when predicting "pain" images. However, we found that, using the same classification model, different XAI methods yielded distinct attribution masks to a given input. This finding raises concerns about the effective use and interpretation of XAI methods, and, more importantly, what regions of the image are truly relevant to the decision-making process of the machine model itself.

Comparing the human and machine facial perception, we observed that the VGG-Face, through the use of the GradCAM technique, presented a clustering feature extraction that is similar to the holistic facial perception of adults. However, further comparisons showed that the facial regions most relevant to the implemented machine models are different to the ones observed by adults. Moreover, there is disagreement between them when analysing neonates during periods of rest. Consequently, these differences suggest that humans and machines can learn with each other in order to improve their current decision-making process of identifying the discriminant information that defines neonatal pain.

As future work, we intend to investigate the agreement between humans and machine to each region of the face, segmenting each one accordingly to clinical scales, obtaining the contour masks of the nasolabial furrow, eyes, mouth, etc. Also, we propose to analyse the contribution of each facial feature to the output of machine models in terms of the confidence score to a given input, since current XAI methods do not account for semantic information yet.

REFERENCES

ANAND, Kanwaljeet JS. Defining pain in newborns: need for a uniform taxonomy? **Acta Pædiatrica**, Wiley Online Library, v. 106, n. 9, p. 1438–1444, 2017.

ANAND, Kanwaljeet JS; CARR, David B. The neuroanatomy, neurophysiology, and neurochemistry of pain, stress, and analgesia in newborns and children. **Pediatric Clinics of North America**, Elsevier, v. 36, n. 4, p. 795–822, 1989.

ANAND, Kanwaljeet JS; HICKEY, Paul R, et al. Pain and its effects in the human neonate and fetus. **N Engl j Med**, v. 317, n. 21, p. 1321–1329, 1987.

BALDA, Rita et al. The recognition of facial expression of pain in full-term newborns by parents and health professionals. **Archives of pediatrics & adolescent medicine**, American Medical Association, v. 154, n. 10, p. 1009–1016, 2000.

BAR, Yaniv et al. Deep learning with non-medical training used for chest pathology identification. In: INTERNATIONAL SOCIETY FOR OPTICS and PHOTONICS. MEDICAL Imaging 2015: Computer-Aided Diagnosis. [S.l.: s.n.], 2015. v. 9414, p. 94140v.

BARROS, Marina Carvalho de Moraes et al. Identification of pain in neonates: the adults' visual perception of neonatal facial features. **Journal of Perinatology**, Nature Publishing Group, v. 41, n. 9, p. 2304–2308, 2021.

BEYER, Judith E et al. Patterns of postoperative analgesic use with adults and children following cardiac surgery. **Pain**, Elsevier, v. 17, n. 1, p. 71–81, 1983.

BOSER, Bernhard E; GUYON, Isabelle M; VAPNIK, Vladimir N. A training algorithm for optimal margin classifiers. In: PROCEEDINGS of the 5th Annual ACM Workshop on Computational Learning Theory. [S.l.: s.n.]. P. 144–152.

BRAHNAM, Sheryl et al. Machine assessment of neonatal facial expressions of acute pain. **Decision Support Systems**, Elsevier, v. 43, n. 4, p. 1242–1254, 2007.

BRAHNAM, Sheryl et al. Machine recognition and representation of neonatal facial displays of acute pain. **Artificial intelligence in medicine**, Elsevier, v. 36, n. 3, p. 211–222, 2006.

BRAHNAM, Sheryl et al. SVM classification of neonatal facial images of pain. **International Workshop on Fuzzy Logic and Applications**, p. 121–128, 2005.

CARLINI, Lucas Pereira et al. A Convolutional Neural Network-based Mobile Application to Bedside Neonatal Pain Assessment. In: IEEE COMPUTER SOCIETY. 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). [S.l.: s.n.], 2021. P. 394–401.

CARLINI, Lucas Pereira et al. A Visual Perception Framework to Analyse Neonatal Pain in Face Images. In: CAMPILHO, Aurélio; KARRAY, Fakhri; WANG, Zhou (Eds.). **Image Analysis and Recognition**. Proceedings of the 17th International Conference on Image Analysis and

Recognition, ICIAR 2020. Cham: Springer International Publishing, 2020. v. 12131. (Lecture Notes in Computer Science), p. 233–243. ISBN 978-3-030-50347-5.

CARLINI, Lucas Pereira et al. Neonatal Pain Scales and Human Visual Perception: An Exploratory Analysis Based on Facial Expression Recognition and Eye-Tracking. In: SPRINGER INTERNATIONAL PUBLISHING. PATTERN Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II. [S.l.: s.n.], 2021. P. 62–76.

CASSIA XAVIER BALDA, Rita de; GUINSBURG, Ruth. Perceptions of neonatal pain. **Neoreviews**, American Academy of Pediatrics, v. 8, n. 12, e533–e542, 2007.

CHATFIELD, Ken et al. Return of the devil in the details: Delving deep into convolutional nets. **arXiv preprint arXiv:1405.3531**, 2014.

CHENG, Xiaoying et al. Artificial Intelligence Based Pain Assessment Technology in Clinical Application of Real-World Neonatal Blood Sampling. **Diagnostics**, MDPI, v. 12, n. 8, p. 1831, 2022.

CORNELISSEN, Laura et al. Postnatal temporal, spatial and modality tuning of nociceptive cutaneous flexion reflexes in human infants. **PloS one**, Public Library of Science San Francisco, USA, v. 8, n. 10, e76470, 2013.

COUTRIN, Gabriel AS et al. Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison. In: PROCEEDINGS of the 3rd International Conference on Medical Imaging and Computer-Aided Diagnosis, MICAD 2022. [S.l.]: Springer LNEE, 2022.

CRAIG, Kenneth D et al. A comparison of two measures of facial activity during pain in the newborn child. **Journal of Pediatric Psychology**, Oxford University Press, v. 19, n. 3, p. 305–318, 1994.

CRISTIANINI, Nello; SHAWE-TAYLOR, John, et al. **An introduction to support vector machines and other kernel-based learning methods**. [S.l.]: Cambridge university press, 2000.

CRUZ, MD; FERNANDES, AM; OLIVEIRA, CR. Epidemiology of painful procedures performed in neonates: a systematic review of observational studies. **European Journal of Pain**, Wiley Online Library, v. 20, n. 4, p. 489–498, 2016.

DARMSTADT, Gary L; SHIFFMAN, Jeremy; LAWN, Joy E. Advancing the newborn and stillbirth global agenda: priorities for the next decade. **Archives of disease in childhood**, BMJ Publishing Group Ltd, v. 100, Suppl 1, s13–s18, 2015.

DE LA TORRE, Jimmy. The generalized DINA model framework. **Psychometrika**, Springer, v. 76, n. 2, p. 179–199, 2011.

DENG, Jiankang et al. **RetinaFace: Single-stage Dense Face Localisation in the Wild**. [S.l.: s.n.], 2019. arXiv: 1905.00641 [cs.CV].

DIATCHENKO, Luda et al. Genetic architecture of human pain perception. **TRENDS in Genetics**, Elsevier, v. 23, n. 12, p. 605–613, 2007.

DIATCHENKO, Luda et al. Idiopathic pain disorders–pathways of vulnerability. **Pain**, LWW, v. 123, n. 3, p. 226–230, 2006.

DOESBURG, Sam M et al. Neonatal pain-related stress, functional cortical activity and visual-perceptual abilities in school-age children born at extremely low gestational age. **PAIN®**, Elsevier, v. 154, n. 10, p. 1946–1952, 2013.

FABRIZI, Lorenzo et al. A shift in sensory processing that enables the developing human brain to discriminate touch from pain. **Current Biology**, Elsevier, v. 21, n. 18, p. 1552–1558, 2011.

FISHER, Ronald A. The use of multiple measurements in taxonomic problems. **Annals of eugenics**, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

FUKUNAGA, Keinosuke. **Introduction to statistical pattern recognition**. [S.l.]: Elsevier, 2013.

GRUNAU, Ruth Eckstein. Personal perspectives: Infant pain—A multidisciplinary journey. **Paediatric and Neonatal Pain**, Wiley Online Library, v. 2, n. 2, p. 50–57, 2020.

GRUNAU, Ruth VE; CRAIG, Kenneth D. Pain expression in neonates: facial action and cry. **Pain**, Elsevier, v. 28, n. 3, p. 395–410, 1987.

GRUNAU, Ruth VE; JOHNSTON, C Celeste; CRAIG, Kenneth D. Neonatal facial and cry responses to invasive and non-invasive procedures. **Pain**, Elsevier, v. 42, n. 3, p. 295–305, 1990.

GUINSBURG, Ruth. Avaliação e tratamento da dor no recém-nascido. **J Pediatr (Rio J)**, v. 75, n. 3, p. 149–60, 1999.

HALL, Mark Andrew. Correlation-based feature selection for machine learning. Citeseer, 1999.

HASLAM, Diana R. Age and the perception of pain. **Psychonomic Science**, Springer, v. 15, n. 2, p. 86–87, 1969.

HEIDERICH, Tatiany M; GUINSBURG, R; LESLIE, Ana T S. **Desenvolvimento de *Software* para Identificar a Expressão Facial de Dor do Recém-Nascido**. 2013. PhD thesis – Universidade Federal de São Paulo (UNIFESP), São Paulo.

HEIDERICH, Tatiany Marcondes; LESLIE, Ana Teresa Figueiredo Stochero; GUINSBURG, Ruth. Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. **Acta Paediatrica**, Wiley Online Library, v. 104, n. 2, e63–e69, 2015.

HINTON, Geoffrey; SRIVASTAVA, Nitish; SWERSKY, Kevin. **Overview of mini-batch gradient descent**. [S.l.: s.n.], 2012. [Online; accessed 23-Julho-2020]. Available from: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HOLMQVIST, Kenneth et al. **Eye tracking: A comprehensive guide to methods and measures**. [S.l.]: OUP Oxford, 2011.

HOLSTI, Liisa et al. Prior pain induces heightened motor responses during clustered care in preterm infants in the NICU. **Early human development**, Elsevier, v. 81, n. 3, p. 293–302, 2005.

HUMMEL, Pat; DIJK, Monique van. Pain assessment: current status and challenges. In: ELSEVIER, 4. SEMINARS in Fetal and Neonatal medicine. [S.l.: s.n.], 2006. v. 11, p. 237–245.

HWU, Tiffany et al. Matching Representations of Explainable Artificial Intelligence and Eye Gaze for Human-Machine Interaction. **arXiv preprint arXiv:2102.00179**, 2021.

JENI, László A; COHN, Jeffrey F; KANADE, Takeo. Dense 3D face alignment from 2D videos in real-time. In: IEEE. 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG). [S.l.: s.n.], 2015. v. 1, p. 1–8.

JOLLIFFE, I.T.; SPRINGER-VERLAG. **Principal Component Analysis**. [S.l.]: Springer, 2002. (Springer Series in Statistics). ISBN 9780387954424. Available from: https://books.google.com.br/books?id=%5C_olByCrhjwIC.

KENDALL, Maurice G. A new measure of rank correlation. **Biometrika**, JSTOR, v. 30, n. 1/2, p. 81–93, 1938.

KIRA, Kenji; RENDELL, Larry A. A practical approach to feature selection. In: MACHINE Learning Proceedings 1992. [S.l.]: Elsevier, 1992. P. 249–256.

KRECHEL, Susan W; BILDNER, Judy. CRIES: a new neonatal postoperative pain measurement score. Initial testing of validity and reliability. **Pediatric Anesthesia**, Wiley Online Library, v. 5, n. 1, p. 53–61, 1995.

LAWRENCE, Jocelyn et al. The development of a tool to assess neonatal pain. **Neonatal network: NN**, v. 12, n. 6, p. 59–66, 1993.

LIN, Tsung-Yu; ROYCHOWDHURY, Aruni; MAJI, Subhransu. Bilinear cnn models for fine-grained visual recognition. In: PROCEEDINGS of the IEEE international conference on computer vision. [S.l.: s.n.], 2015. P. 1449–1457.

MACQUEEN, James et al. Some methods for classification and analysis of multivariate observations. In: 14. PROCEEDINGS of the fifth Berkeley symposium on mathematical statistics and probability. [S.l.: s.n.], 1967. v. 1, p. 281–297.

MANN, Henry B; WHITNEY, Donald R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, JSTOR, p. 50–60, 1947.

MARTÍN ABADI et al. **TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems**. [S.l.: s.n.], 2015. Software available from tensorflow.org. Available from: https://www.tensorflow.org/.

MERSKEY, H. On the development of pain. **Headache: The Journal of Head and Face Pain**, Wiley Online Library, v. 10, n. 3, p. 116–123, 1970.

MIKOŁAJCZYK, Agnieszka; GROCHOWSKI, Michał. Data augmentation for improving deep learning in image classification problem. In: IEEE. 2018 international interdisciplinary PhD workshop (IIPhDW). [S.l.: s.n.], 2018. P. 117–122.

OSTER, Harriet. Facial expression and affect development. In: THE development of affect. [S.l.]: Springer, 1978. P. 43–75.

OWENS, Mark E. Pain in infancy: conceptual and methodological issues. **Pain**, Elsevier, v. 20, n. 3, p. 213–230, 1984.

P., Ekman.; FRIESEN, W.V. **Manual for the Facial Action Coding System**. 1. ed. Palo Alto. CA: Consulting Psychologists Press., 1978.

PARKHI, Omkar M.; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep Face Recognition. In: BRITISH Machine Vision Conference. [S.l.: s.n.], 2015.

PAUL, Rahul et al. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. **Tomography**, Grapho Publications, v. 2, n. 4, p. 388, 2016.

PEREZ, Luis; WANG, Jason. The effectiveness of data augmentation in image classification using deep learning. **arXiv preprint arXiv:1712.04621**, 2017.

RASSADIN, Alexandr; GRUZDEV, Alexey; SAVCHENKO, Andrey. Group-level emotion recognition using transfer learning from face identification. In: PROCEEDINGS of the 19th ACM International Conference on Multimodal Interaction. [S.l.: s.n.], 2017. P. 544–548.

RAZALI, Nornadiah Mohd; WAH, Yap Bee, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. **Journal of statistical modeling and analytics**, v. 2, n. 1, p. 21–33, 2011.

SALEKIN, Md Sirajus et al. First investigation into the use of deep learning for continuous assessment of neonatal postoperative pain. In: IEEE. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). [S.l.: s.n.], 2020. P. 415–419.

SALEKIN, Md Sirajus et al. Future roles of artificial intelligence in early pain management of newborns. **Paediatric and Neonatal Pain**, Wiley Online Library, v. 3, n. 3, p. 134–145, 2021.

SALEKIN, Md Sirajus et al. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. **Computers in Biology and Medicine**, Elsevier, v. 129, p. 104150, 2021.

SALEKIN, Md Sirajus et al. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. **Computers in Biology and Medicine**, Elsevier, v. 129, p. 104150, 2021.

SARGANO, Allah Bux et al. Human action recognition using transfer learning with deep representations. In: IEEE. 2017 International joint conference on neural networks (IJCNN). [S.l.: s.n.], 2017. P. 463–469.

SASAKI, Yutaka et al. The truth of the F-measure. **Teach tutor mater**, v. 1, n. 5, p. 1–5, 2007.

SCHILLER, Dominik et al. Relevance-based data masking: a model-agnostic transfer learning approach for facial expression recognition. **Frontiers in Computer Science**, Frontiers, v. 2, p. 6, 2020.

SELVARAJU, Ramprasaath R et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: PROCEEDINGS of the IEEE international conference on computer vision. [S.l.: s.n.], 2017. P. 618–626.

SEXTON, Randall S; DORSEY, Robert E; SIKANDER, Naheel A. Simultaneous optimization of neural network function and architecture algorithm. **Decision Support Systems**, Elsevier, v. 36, n. 3, p. 283–296, 2004.

SHAPHIRO, S; WILK, M. An analysis of variance test for normality. **Biometrika**, v. 52, n. 3, p. 591–611, 1965.

SILVA, Giselle Valério Teixeira da et al. What Facial Features Does the Pediatrician Look to Decide That a Newborn Is Feeling Pain? **American Journal of Perinatology**, Thieme Medical Publishers, Inc., 2021.

SIMONS, Sinno HP et al. Do we still hurt newborn babies?: A prospective study of procedural pain and analgesia in neonates. **Archives of pediatrics & adolescent medicine**, American Medical Association, v. 157, n. 11, p. 1058–1064, 2003.

SLATER, Rebeccah et al. Premature infants display increased noxious-evoked neuronal activity in the brain compared to healthy age-matched term-born infants. **Neuroimage**, Elsevier, v. 52, n. 2, p. 583–589, 2010.

SOARES, Juliana do Carmo Azevedo et al. Looking at neonatal facial features of pain: do health and non-health professionals differ? **Jornal de pediatria**, Elsevier, 2021.

SOUTHALL, David P et al. Invasive procedures in children receiving intensive care. **BMJ: British Medical Journal**, BMJ Publishing Group, v. 306, n. 6891, p. 1512, 1993.

STEVENS, Bonnie et al. Premature infant pain profile: development and initial validation. **The Clinical journal of pain**, LWW, v. 12, n. 1, p. 13–22, 1996.

STUDENT. The probable error of a mean. **Biometrika**, JSTOR, p. 1–25, 1908.

SUNDARARAJAN, Mukund; TALY, Ankur; YAN, Qiqi. Axiomatic attribution for deep networks. **arXiv preprint arXiv:1703.01365**, 2017.

TAMANAKA, Fernanda G. et al. Neonatal pain assessment: A Kendall analysis between clinical and visually perceived facial features. **Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization**, Taylor & Francis, v. 0, n. 0, p. 1–10, 2022. DOI: 10.1080/21681163.2022.2044909. eprint: https://doi.org/10.1080/21681163.2022.2044909. Available from: https://doi.org/10.1080/21681163.2022.2044909.

TERUEL, G. F. et al. Analysis and recognition of pain in 2D face images of full term and healthy newborns. **Proceedings of the XV Encontro Nacional de Inteligencia Artificial, ENIAC 2018**, p. 228–239, 2018.

TERUEL, Gilberto F; THOMAZ, Carlos E. **Análise e Reconhecimento de Dor em Imagens 2D Frontais de Recém-Nascidos a Termo e Saudáveis**. 2018. MA thesis – Centro Universitário FEI (FEI), São Bernardo do Campo, SP.

THOMAZ, Carlos E; KITANI, Edson Caoru; GILLIES, Duncan Fyfe. A maximum uncertainty LDA-based approach for limited sample size problems—with application to face recognition. **Journal of the Brazilian Computer Society**, Springer, v. 12, n. 2, p. 7–18, 2006.

THOMAZ, Carlos E et al. A multivariate statistical analysis of the developing human brain in preterm infants. **Image and Vision Computing**, Elsevier, v. 25, n. 6, p. 981–994, 2007.

VELDEN, Bas HM van der et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. **arXiv preprint arXiv:2107.10912**, 2021.

VINALL, Jillian et al. Invasive procedures in preterm children: brain and cognitive development at school age. **Pediatrics**, Am Acad Pediatrics, v. 133, n. 3, p. 412–421, 2014.

WILLIAMS, Morika D; LASCELLES, B Duncan X. Early neonatal pain—A review of clinical and experimental implications on painful conditions later in life. **Frontiers in pediatrics**, Frontiers, p. 30, 2020.

WOODROW, Kenneth M et al. Pain tolerance: differences according to age, sex and race. **Psychosomatic medicine**, LWW, v. 34, n. 6, p. 548–556, 1972.

ZAMZMI, Ghada et al. Neonatal pain expression recognition using transfer learning. **arXiv preprint arXiv:1807.01631**, 2018.

ZAMZMI, Ghada et al. Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN). In: IEEE. 2019 International Joint Conference on Neural Networks (IJCNN). [S.l.: s.n.], 2019. P. 1–7.

ZHOU, Bolei et al. Learning deep features for discriminative localization. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016. P. 2921–2929.

APPENDIX A – FINAL VGG-FACE ARCHITECTURE.

	Type of layer	Number of filters	Feature Map Size	Filter Size	Strides	Padding
	Image input		224 x 224 x 3			
Group 1	1º convolutional layer ReLU	64	224 x 224 x 64 224 x 224 x 64	3 x 3	1 x 1	1 x 1
	2º convolutional layer ReLU	64	224 x 224 x 64 224 x 224 x 64	3 x 3	1 x 1	1 x 1
	MaxPooling	1	112 x 112 x 64	2 x 2	2 x 2	0 x 0
Group 2	3º convolutional layer ReLU	128	112 x 112 x 128 112 x 112 x 128	3 x 3	1 x 1	1 x 1
	4º convolutional layer ReLU	128	112 x 112 x 128 112 x 112 x 128	3 x 3	1 x 1	1 x 1
	MaxPooling	1	56 x 56 x 128	2 x 2	2 x 2	0 x 0
Group 3	5º convolutional layer ReLU	256	56 x 56 x 256 56 x 56 x 256	3 x 3	1 x 1	1 x 1
	6º convolutional layer ReLU	256	56 x 56 x 256 56 x 56 x 256	3 x 3	1 x 1	1 x 1
	7º convolutional layer ReLU	256	56 x 56 x 256 56 x 56 x 256	3 x 3	1 x 1	1 x 1
	MaxPooling	1	28 x 28 x 256	2 x 2	2 x 2	0 x 0
Group 4	8º convolutional layer ReLU	512	28 x 28 x 512 28 x 28 x 512	3 x 3	1 x 1	1 x 1
	9º convolutional layer ReLU	512	28 x 28 x 512 28 x 28 x 512	3 x 3	1 x 1	1 x 1
	10º convolutional layer ReLU	512	28 x 28 x 512 28 x 28 x 512	3 x 3	1 x 1	1 x 1
	MaxPooling	1	14 x 14 x 512	2 x 2	2 x 2	0 x 0
Group 5	11º convolutional layer ReLU	512	14 x 14 x 512 14 x 14 x 512	3 x 3	1 x 1	1 x 1
	12º convolutional layer ReLU	512	14 x 14 x 512 14 x 14 x 512	3 x 3	1 x 1	1 x 1
	13º convolutional layer ReLU	512	14 x 14 x 512 14 x 14 x 512	3 x 3	1 x 1	1 x 1
	MaxPooling	1	7 x 7 x 512	2 x 2	2 x 2	0 x 0
	1º fully connected layer ReLU DropOut		512 x 1 512 x 1 512 x 1			
	2º fully connected layer ReLU DropOut		512 x 1 512 x 1 512 x 1			
	Dense layer		2 x 1			
	Output		2 x 1			

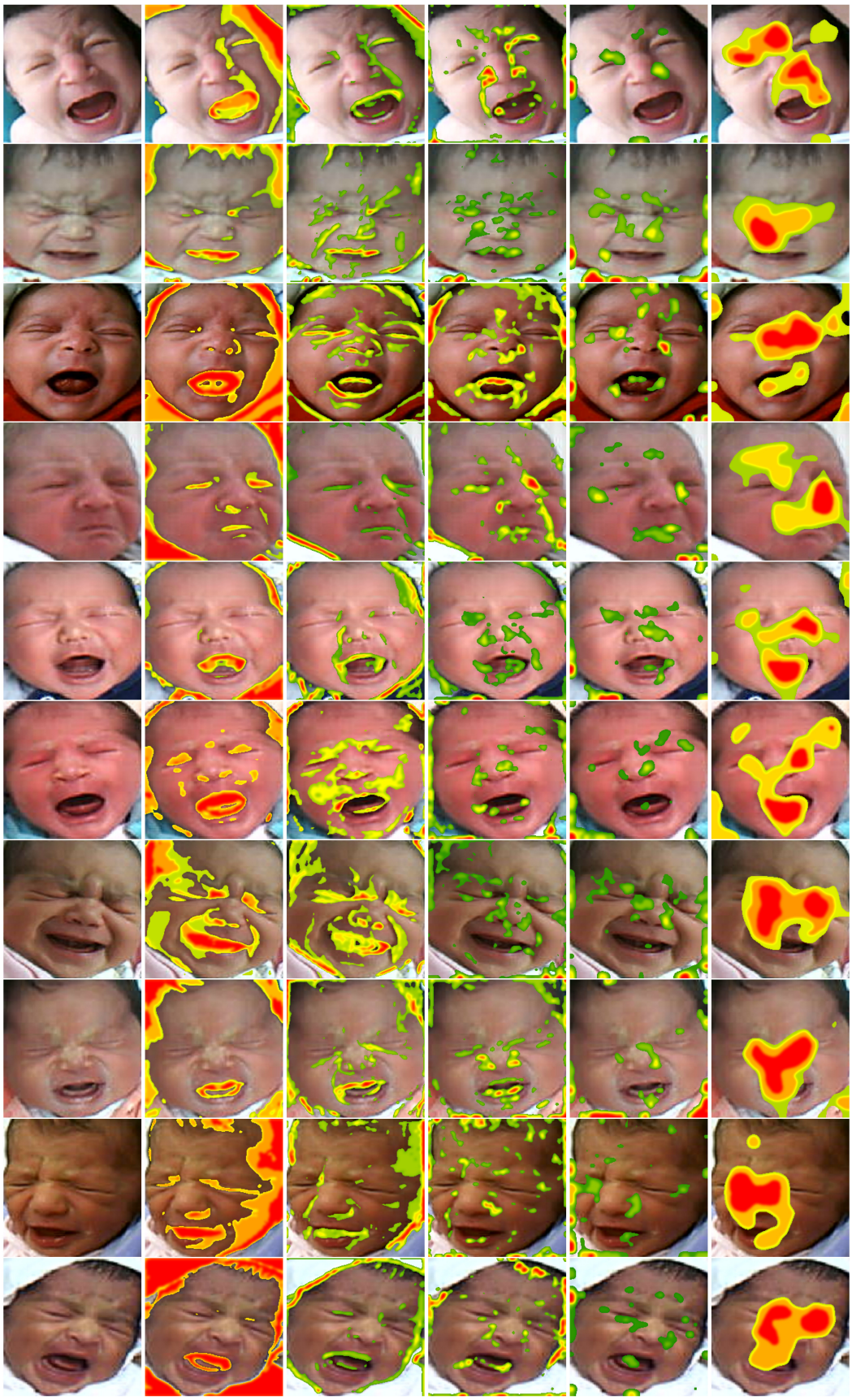
Obs: height x width x channel.

APPENDIX B – ADAPTED N-CNN ARCHITECTURE.

Type of Layer		Number of Filters	Feature Map Size	Filter Size	Strides	Padding
Image input			120 x 120 x 3			
Left Branch	MaxPooling	1	12 x 12 x 3	10 x 10	10 x 10	0 x 0
Central Branch	1º convolutional layer	64	120 x 120 x 64	5 x 5	1 x 1	0 x 0
	Leaky ReLU		120 x 120 x 64			
	MaxPooling	1	40 x 40 x 64	3 x 3	3 x 3	0 x 0
	2º convolutional layer	64	40 x 40 x 64	2 x 2	1 x 1	0 x 0
	Leaky ReLU		40 x 40 x 64			
	MaxPooling	1	14 x 14 x 64	3 x 3	3 x 3	0 x 0
	Resizing		12 x 12 x 64			
	DropOut		12 x 12 x 64			
Right Branch	Convolutional layer	64	120 x 120 x 64	5 x 5	1 x 1	0 x 0
	Leaky ReLU		120 x 120 x 64			
	MaxPooling	1	12 x 12 x 64	10 x 10	10 x 10	0 x 0
	DropOut		12 x 12 x 64			
Merging Branches			12 x 12 x 131			
Convolutional layer		64	12 x 12 x 64	2 x 2	1 x 1	0 x 0
ReLU			12 x 12 x 64			
MaxPooling		1	6 x 6 x 64	2 x 2	2 x 2	0 x 0
Fully connected layer			8 x 1			
ReLU			8 x 1			
DropOut			8 x 1			
Dense Layer			2 x 1			
Output			2 x 1			

Obs: height x width x channel.

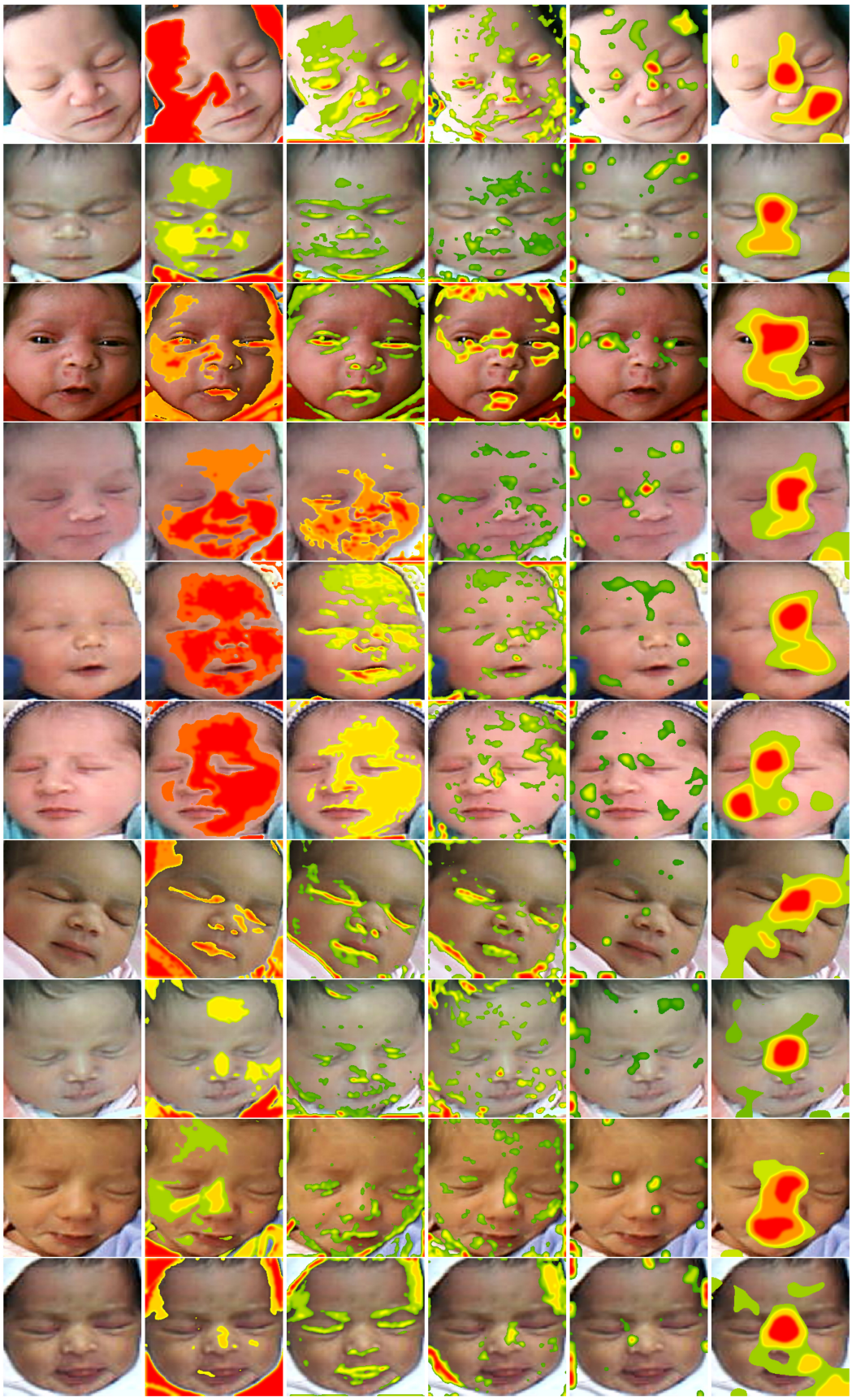
APPENDIX C – VGG-FACE'S LAYERS DEPTH IMPACT ON GRADCAM FEATURE EXTRACTION ON "PAIN" IMAGES.



From left to right: input image, 1st convolutional group, 2nd group, 3rd group, 4th group, and 5th group.

Source: Author, using the UNIFESP Face Image Dataset.

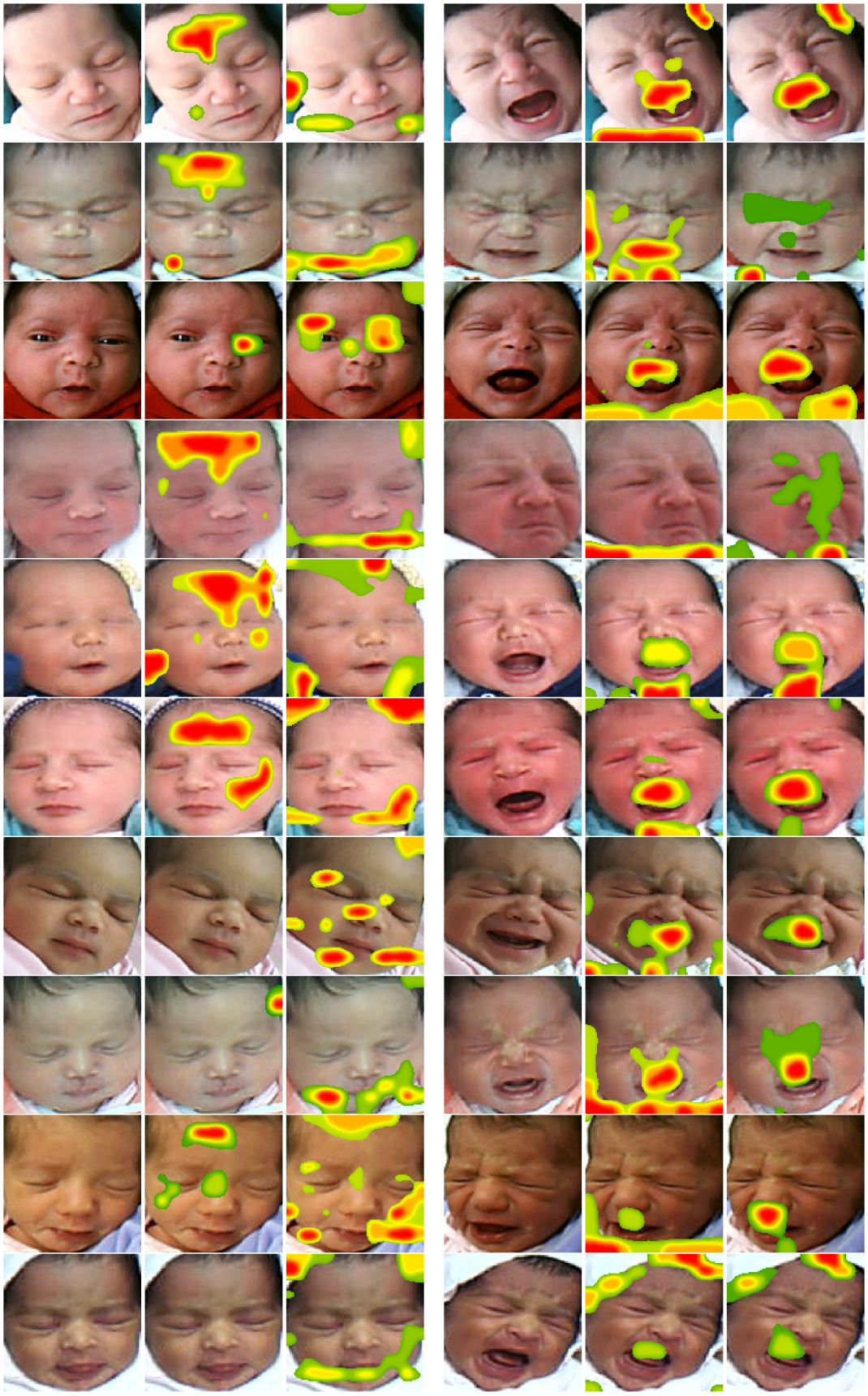
APPENDIX D – VGG-FACE'S LAYERS DEPTH IMPACT ON GRADCAM FEATURE EXTRACTION ON "NO PAIN" IMAGES.



From left to right: input image, 1st convolutional group, 2nd group, 3rd group, 4th group, and 5th group.

Source: Author, using the UNIFESP Face Image Dataset.

APPENDIX E – N-CNN'S LAYERS DEPTH IMPACT ON GRADCAM FEATURE EXTRACTION ON "NO PAIN" AND "PAIN" IMAGES.



From left to right: “no pain” input image, merge layer, final convolutional layer, “pain” input image, merge layer, and final convolutional layer.

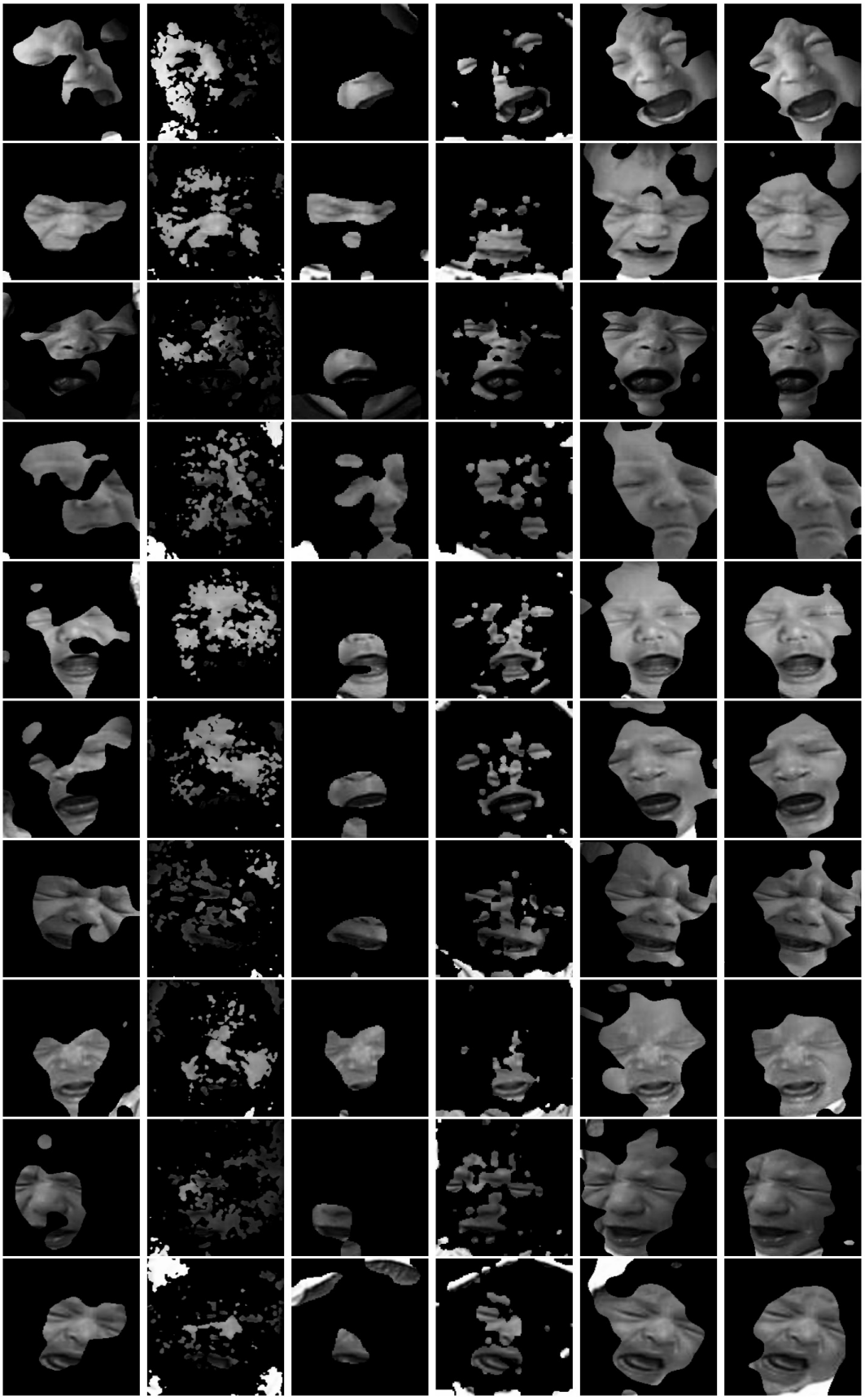
Source: Author, using the UNIFESP Face Image Dataset.

APPENDIX F – THE HEATMAPS OF FACIAL FEATURES EXTRACTED BY EACH EXTRACTION MODEL WHEN ASSESSING “PAIN” IMAGES.

From left to right: VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts.

Source: Author, using the UNIFESP Face Image Dataset.

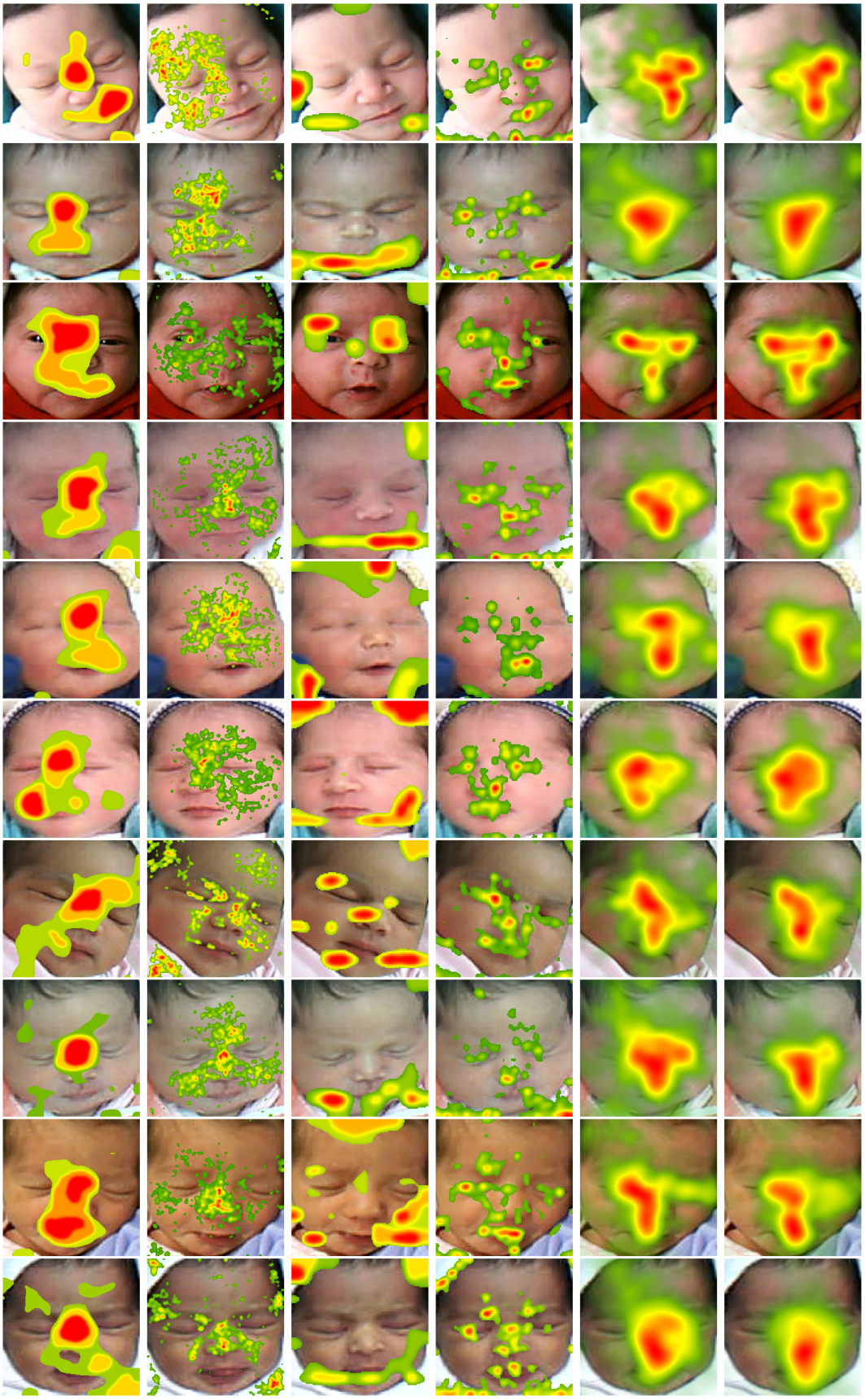
APPENDIX G – THE PERCEIVED FACIAL AREAS BY EACH EXTRACTION MODEL WHEN ASSESSING “PAIN” IMAGES.



From left to right: VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts.

Source: Author, using the UNIFESP Face Image Dataset.

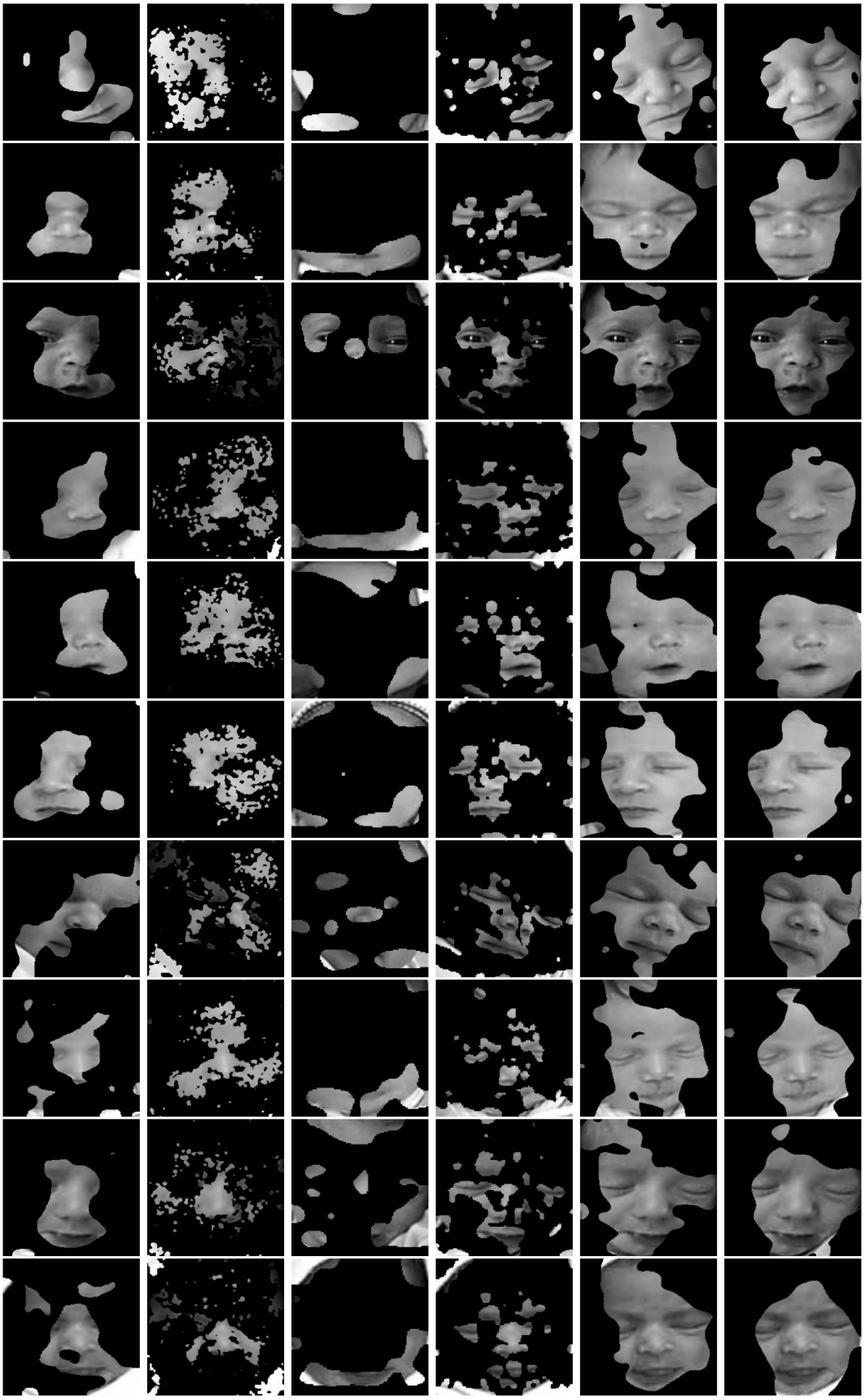
APPENDIX H – THE HEATMAPS OF FACIAL FEATURES EXTRACTED BY EACH EXTRACTION MODEL WHEN ASSESSING “NO PAIN” IMAGES.



From left to right: VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts.

Source: Author, using the UNIFESP Face Image Dataset.

APPENDIX I – THE PERCEIVED FACIAL AREAS BY EACH EXTRACTION MODEL WHEN ASSESSING “NO PAIN” IMAGES.



From left to right: VGG-Face's (1) GradCAM and (2) Integrated Gradients, N-CNN's (3) GradCAM and (4) Integrated Gradients, (5) Parents, and (6) Experts.

Source: Author, using the UNIFESP Face Image Dataset.