APRENDIZADO PROFUNDO PARA RECONHECIMENTO DE EXPRESSÕES FACIAIS DE DOR EM FETOS HUMANOS

GUILHERME MENDES MARQUES DE OLIVEIRA

APRENDIZADO PROFUNDO PARA RECONHECIMENTO DE EXPRESSÕES FACIAIS DE DOR EM FETOS HUMANOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Adriano Alonso Veloso
Coorientador: Nivio Ziviani

Belo Horizonte, Minas Gerais
Fevereiro de 2022

GUILHERME MENDES MARQUES DE OLIVEIRA

A DEEP LEARNING MODEL FOR AUTOMATIC RECOGNITION OF PAIN FACIAL EXPRESSIONS ON HUMAN FETUSES

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Adriano Alonso Veloso
Co-Advisor: Nivio Ziviani

Belo Horizonte, Minas Gerais
February 2022

Oliveira, Guilherme Mendes Marques de

D1234p A DEEP LEARNING MODEL FOR AUTOMATIC RECOGNITION OF PAIN FACIAL EXPRESSIONS ON HUMAN FETUSES / Guilherme Mendes Marques de Oliveira. — Belo Horizonte, Minas Gerais, 2022

xxii, 68 f. : il. ; 29cm

Dissertação (mestrado) — Federal University of Minas Gerais

Orientador: Adriano Alonso Veloso

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Acknowledgments

I'm deeply grateful to both my advisor Prof. Adriano Veloso and co-advisor Prof. Nivio Ziviani, they taught me plenty in the scientific scope and further beyond, in addition to their invaluable advices, continuous support, and patience. I would also like to express my gratitude for the Fetal Pain Study Group from Universidade de São Paulo for their profitable assistance and insights, specially to Prof. Daniel Ciampi de Andrade and Prof. Lisandra Bernardes.

"Every brilliant experiment, like every great work of art, starts with an act of imagination [...]."

(Jonah Lehrer)

Resumo

A base neurobiológica estrutural do cérebro necessária para a integração do que chamamos de dor está presente no feto humano após a 20ª semana de gestação e presume-se que no terceiro trimestre o feto humano pode sentir dor aguda e provavelmente crônica. A dor está associada não apenas ao sofrimento, mas também a uma maior incidência de declínio cognitivo, portanto, a avaliação do comportamento relacionado à dor fetal pode melhorar drasticamente a sensibilidade e a especificidade dos marcadores de vitalidade e bem-estar fetais usados atualmente. No entanto, a avaliação dos comportamentos relacionados à dor é relativamente limitada dentro do útero, pois a ocorrência de movimento fetal intrauterino não é uma indicação de que o feto está sentindo dor. A estratégia de usar a expressão facial como marcador substituto da presença de dor já foi utilizada em outros cenários, como em recém-nascidos, em idosos não comunicantes, em adultos com dificuldade de fala e em outros mamíferos não primatas. Ao registrar as expressões faciais de fetos submetidos à cirurgia intrauterina, utilizamos a reação de dor aguda desencadeada pela injeção de anestésico administrada antes do ato cirúrgico como um modelo confiável e padronizado de dor aguda. Como o rosto do feto se assemelha ao rosto do adulto em muitos aspectos, empregamos uma rede neural profunda de reconhecimento de rosto treinada em milhões de rostos de adultos e ajustamos seus parâmetros para detectar sinais de dor nas expressões faciais do feto. Mostramos que a dor fetal pode ser detectada com índice de acertos e AUC ROC de aproximadamente 88% e 99%, respectivamente.

Palavras-chave: Feto Humano, Dor, Dor Fetal, Ultrassom 4D, Aprendizado Profundo, Visão Computacional.

Abstract

The neurobiological structural brain background necessary for the integration of what we call pain is present in the human fetus after the 20th week of gestation and it is assumed that in the third trimester of gestation the human fetus can experience acute and likely chronic pain. Pain is associated not only with physical suffering but also with a higher incidence of cognitive decline, thus fetal pain-related behavior evaluation may drastically improve sensitivity and specificity of currently used fetal vitality and wellbeing markers. However, assessment of pain-related behaviors is relatively limited inside the uterus, as the occurrence of intrauterine fetal movement is not an indication that the fetus is feeling pain. The strategy of using facial expression as a surrogate marker of the presence of pain has already been used in other scenarios, such as in newborns, in non-communicating elderlies, in speech-impaired adults, and in other non-primate mammals. By recording facial expressions of fetuses undergoing intrauterine surgery, we used the acute pain reaction triggered by the anesthetic injection administered prior to surgical act as a reliable and standardized model of acute pain. Since the fetus face resembles the adult face in many aspects, we employ a face recognition deep neural network trained on millions of adult faces, and tuned its parameters in order to detect pain signals in fetus facial expressions. We show that fetal pain can be detected with an accuracy and ROC AUC of aproximadetedly 88% and 99%, respectively.

Palavras-chave: Human Fetus, Pain, Fetal Pain, 4D Ultrasound, Deep Learning, Computer Vision.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Fetal pain-related behavior evaluation may drastically improve sensitivity and specificity of currently used fetal vitality and wellbeing markers, allowing for better detection of fetal distress with less false positive rates. We propose to (i) create a new protocol for assessing signals of pain in fetal facial expressions using a sophisticated deep learning model, (ii) inaugurate the possibility to integrate this automated protocol in 4D Ultrasound (4D-US) machines used in the everyday evaluation of fetal vitality and (iii) advance the studies towards fetal health and vitality estimation still in the prenatal phase.

Chronic pain affects 20% of the general population and is associated not only with physical suffering but with a higher incidence of cognitive decline and worse performance in adult life. The neurobiological structural brain background necessary for the integration of what we call pain is present in the human fetus after the 20th week of gestation and it is assumed that in the third trimester of gestation the human fetus can experience acute and likely chronic pain.

The goal of the present work is to determine whether human fetuses demonstrate discriminative acute behavioral responses to nociceptive input that can be automatically detected using deep learning applied on 4D-US images of fetal faces. Additionally, creating datasets to enable deep learning applications on the subject, furthermore, the greater goal is to assess general fetal vitality, however, the preliminary steps include assessing fetal pain.

Machine learning is a form of AI (Artificial Intelligence) which completely changed the paradigm of traditional computer science algorithms: it is capable of consuming data as input, mapping each sample to its corresponding answer. In this case, the input data are 4D-US images of fetuses, where those images have been labeled by medical professionals. Therefore, this project owns the adequate arrangement to explore the application of machine learning to assess fetuses. Further, AI has developed to the point of treating novel problems in the computer vision domain that previously could only be poorly assessed. Accordingly, the usage of the novel CNNs (Convolutional Neural Networks) has the potential to lead to great performance across many different vision applications.

1.1 Motivation

Diagnosing newborns is significantly more challenging than diagnosing adults since they are unable to self-report pain. As a consequence, to better recognize pain, specialists have coupled nonverbal feedback such as facial expressions with physiological data. This approach has been evaluated and proven to comprise a set of reliable pain markers leading to the creation of multiple pain scales like the NFCS (Neonatal Facial Coding System by Grunau et al.) for neonates or it's adaptation for fetuses in the third trimester of gestation, the fetal-5 scale. Prolonged exposure to painful states can have a variety of negative repercussions, including psychological problems, being especially dangerous in neonates and fetuses. Early detection of painful fetal scenarios are really important to enable early interventions and increase the overall fetal wellbeing. Therefore, pain-related behaviors and suffering would request not only specific analgesic treatment during intrauterine procedures and surgeries but also a follow-up to assess the efficacy of such analgesic treatments would be mandatory, something that is not currently done.

Additionally, abortion is another major issue for which, particularly in unintended pregnancies, it has been very common throughout the world between 2015 and 2019 [Bearak et al., 2020]. In 2016, the state of Utah in the United States passed an abortion law requiring anesthesia to women having abortions at 20 weeks of pregnancy or later, however, the anesthesia is not intended for the mother but for the fetus(es) [Fantz, 2016].

Furthermore, through the usage of CNNs (Convolutional Neural Networks), the field of machine learning has reached the point of automatically extracting patterns from the input images, as our experiments show, thus the machine learns by itself, associated with the human defined labels. That's extremely relevant, since not only humans are capable of teaching the machines, but also humans can learn from the machines, creating a virtuous discovery cycle for both ends, as long as the associated explainability is successfully explored. Through the usage of CNNs, extracting patterns of an arbitrary set of images became a powerful task, since artifacts which are both humanly visible and invisible may be detected and identified. Additionally, such artifacts might also corroborate or refute human made hypotheses and statements.

The present project uses deep learning models consuming as input 4D-US images to assess whether a fetus is feeling pain, however, the ultimate objective is to develop a similar application to infer the vitality score a fetus would be given at delivery time. The inference would be achieved based on a regression task, expressing the fetal vitality as the apgar score [Apgar, 1966]. Furthermore, hundreds of fetal videos were collected using 4D-US machines, and those fetuses have been assessed after delivery to obtain their respective apgar scores, creating a large repository of data which ultimately enables the production of large datasets for deep learning applications on the fetal assessment theme.

1.2 Thesis Statement

Since the fetus face resembles the adult face in many aspects, we employ a face recognition deep neural network pre-trained on millions of adult faces, and tuned its parameters in order to detect pain based on the fetal facial expressions present in 4D-US images. We have used modern deep learning techniques such as data augmentation and transfer learning to improve the obtained results, and we show that fetal pain can be detected with an accuracy of roughly 88% and ROC AUC of roughly 99%.

1.3 Contributions

Since fetuses present responses to nociceptive stimuli that can be visually detected using 4D-US images, as will become more clear later on, the present work provides multiple contributions:

- A sophisticated machine learning model that can be used to assist detecting the presence of pain in fetuses;

- Fetal pain datasets cross validation ready to train machine learning models in addition to the methodology to generate them. Further, adaptations of the presented methodology might be employed to generate datasets for a variety of fetal related applications using AI;

- The inauguration of the possibility to integrate automated fetal pain assessment on 4D-US machines, including the automatic generation of associated XAI (eXplainable Artificial Intelligence) elements. Furthermore, the detection of pain-related behavior in the human fetus during healthy and pathological pregnancies is associated to the point that pain-related behaviors and suffering would request not only specific analgesic treatment during intrauterine procedures and surgeries but also a follow-up to assess the efficacy of such analgesic treatments would be mandatory, something that is not currently done. Ultimately, doctors and caregivers would be empowered to detect and understand fetal discomfort related to nociceptive stimulation;

- Opening the way to similarly assess general fetal wellbeing using AI;
- Exploring the explanatory slope provided by machine learning to detect evidence for the presence or absence of pain in fetuses, in addition to crossing the generated explanations with the ones human made as an attempt to corroborate or refute them;

1.4 Organization

Further ahead, this dissertation is organized in the following manner. Chapter 2 presents the most relevant research findings and how they are related to the present work; Chapter 3 provides fundamental knowledge on pregnancy and fetal health which is required to further understand the present work; Chapter 4 provides fundamental knowledge on deep learning which is also required to further understand the present work; Chapter 5 describes the initial phase of the project's methodology, including the works of Bernardes et al. [2018] and Bernardes et al. [2021] which were sources of fundamentals and inspiration for the present work, in addition to original contributions; Chapter 6 describes the final phase of this project's methodology; Chapter 7 presents the the conclusions; and Chapter 8 presents future work.

Chapter 2

Related Work

This chapter is focused on presenting the most relevant research findings and how they are related to the present work.

2.1 General Pain Studies

Presently the IASP (International Association for the Study of Pain) describes pain as an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage [Merskey and Bogduk, 1994]. Additionally, this latest definition has a major difference to the previous one from 1979 stating that verbal description is only one of several behaviors to express pain; inability to communicate does not negate the possibility that a human or a nonhuman animal experiences pain. The new terminology superseded the previous one, which used to classify pain based on a person's capacity to explain the experience, ultimately contemplating an important fundamental of the present work: the qualification of human fetuses to experience pain.

Despite assessing pain on fetuses being a very difficult task many decades ago, with the advent of 4D-US machines, fetuses may easily be recorded in video, even though they are inside the mother's womb. Ultimately, the 4D-US technology enabled far more than assessing just pain, including the appraisal of neurodevelopment, body structure, facial expressions, among more.

2.2 Pain Indicators and Scales for Infants

Due to the incapacity to verbally describe experiencing pain, experts have coupled non-verbal responses from newborns, non-communicating elderlies, speech-impaired adults, and other non-primate mammals. Particularly for neonates, scientists have used feedback such as facial expressions, body movements, crying sounds, changing in behavioral states, psychological indicators and biological markers [Bellieni, 2012] not only as methods to detect pain but, in some cases, also to quantify it.

Facial expressions comprises several features such as eyes squeezing shut or mouth stretching, being a set of important indicators of pain in newborns. The trait has been used by several pain scales, including the NFCS (Neonatal Facial Coding System) [Grunau et al., 1998], however, facial expressions have been used even beyond the detection of pain, particularly as an indicator of fetal brain function [AboEllail and Hata, 2017]. Notwithstanding, body movements such as the activity of arms and legs, or clenching of fists or toes are also used as pain indicators in neonates [Zimmermann, 1991]. Crying is another response provided by newborns, which could indicate pain [Barr et al., 2000] in addition to other non-painful stimuli such as hunger or anger. Also, there are changes in behavioral state that have been proven useful in the assessment of pain including sudden awakening, or crying. Further, the most often utilized physiological markers of pain include changes in heart rate, blood pressure, oxygen saturation, and breathing patterns [Finley and McGrath, 1998]. Finally, biological markers of pain include stress hormones such as cortisol, adrenaline, and beta-endorphins [Herrington et al., 2004].

These pain indicators may be combined to generate unique and understandable pain assessment instruments, known as pain scales, that are used to rate the severity of the discomfort. The sole occurrence of a potential pain related event in neonates are generally not enough to offer guarantees that the cause is truly due to nociceptive stimuli, however, the context of the responses, in addition with the combination of multiple feedback items, can help distinguishing the sources of these symptoms more reliably [Bellieni, 2012]. Further, there exists multiple scales particularly aimed at infants such as Neonatal Infant Pain Scale (NIPS) [Lawrence et al., 1993]; Premature Infant Pain Profile (PIPP) [Stevens et al., 1996]; NFCS [Grunau et al., 1998]; Echelle Douleur Inconfort Nouveau-Né (EDIN) [Debillon et al., 2001]; Cry, Requires O_2, Increased vital signs, Expression, Sleeplessness (CRIES) [Suraseranivongse et al., 2006]; among more.

2.3 On the Feasibility to Assess Fetal Facial Expressions as Pain Markers

In the work of Bernardes et al. [2018], they have reported an experimental model of acute pain in fetuses receiving intrauterine anaesthesia. As there is an evident nociceptive stimulation in this scenario, facial expressions recorded before the intramuscular injection were compared to those recorded immediately after the treatment. They have successfully employed the NFCS [Grunau et al., 1998] pain scale in a randomized and blinded assessment of fetuses, where results have shown discrimination between the resting group and the group subject to the acute pin prick. Finally, they have suggested that when completely and formally confirmed, the NFCS pain scale might allow for the monitoring of analgesic medication during fetal operations, as well as a deeper understanding of the presence of pain behaviors in fetuses with long-term conditions, ultimately improving fetal wellbeing. The follow up work by Bernardes et al. [2021], lead to the creation of the novel fetal-5 pain scale as an adaptation of the NFCS for fetuses, figuring another related work which is also fundamental for the present one, and it is further discussed in Section 3.5.

2.4 Previous Machine Learning Models Applied on Pain Assessment

Machine learning has already been employed on pain assessment, it has been a popular topic of research lately. Particularly for adults, Mauricio et al. [2019] have explored spatiotemporal features extracted from video sequences considering pain stimuli as references in the temporal analysis. Further, for neonates, Zamzmi et al. [2016] presents a multimodal method for automated pain evaluation, in which a pain score is generated by combining a few indicators such as facial expressions, body movements, and changes in vital signs. Subsequently, focusing neonates as well, another research also included crying noises [Zamzmi et al., 2017], moreover, Zamzmi et al. [2018, 2019] have employed CNN based deep neural networks along with data augmentation and transfer learning in order to improve performance, similarly to strategies adopted by the present work.

Finally, de Oliveira et al. [2019] master's dissertation was advised and co-advised by the same researchers as the present work on the same theme, in addition to having access the same raw fetal data. However, de Oliveira et al. has achieved preliminary results suggesting potential. Differently from the first work, the present one has a major methodological improvement. Since fetuses subject to the painful stimulation were simultaneously under the influence of an anesthetic drug, this work restricted the collection and labeling of painful fetal images to a 10 seconds time window right after the anesthesia, discarding all samples that could potentially be biased, as further described in Chapter 5, comprising a major difference from de Oliveira et al. [2019]. Further, despite any similarities, each work built its own datasets along with its particular machine learning pipeline. Moreover, the same strategy was utilized as an attempt to distinguish between actual acute pain and a non-painful disturbing stimulation, as the non-painful group included samples from this later scenario.

Chapter 3

Background on Pregnancy and Fetal Health

This chapter presents fundamental background knowledge in persuance of enabling readers to further understand the relevance of this work and the nature of issues associated with the health of pregancy and fetuses, in addition to the assessment of their context.

3.1 Fetal Signs Suggesting They are Capable of Experiencing Pain

In the present work, all assessed fetuses were in the third gestational trimester i.e., 31.1±2.8 weeks into the pregnancy, when the basic circuitry responsible for nociceptive experiences is believed to be completely functional. That's because, as time passes, fetuses develop their body, contemplating new functionalities according to Table 3.1.

When compared to preterm neonates of similar age, various local environmental factors may impact the perception and experience of nociceptive stimuli in the growing fetus. These environmental factors, as well as the start of postnatal neuronal and behavioral development, establish settings in which pain arises from nociception [Slater et al., 2010], awareness, and past experiences.

Table 3.1. Development of neurological characteristics in fetuses which suggest they are, indeed, capable of feeling pain. Information obtained from Bernardes et al. [2021].

Approximate gestational maturity (weeks)	Associated fetal development phenomenon or phenomena
5	the spinal cord has evolved enough to produce its first synapses [Okado, 1981];
Between 12 and 15	occurrence of thalamic projections from the thalamus to the developing cortex;
Between 23 and 24	within the cortical plate, major corticocortical, thalamocortical, and basal forebrain bundles form synapses [Glover and Fisk, 1999], in addition to free nerve endings and their projections penetrate the spinal cord and completely develop [Fitzgerald, 1987];
25	brain blood flow, noradrenaline release [Giannakoulopoulos et al., 1999], and behavioral changes caused by noxious stimuli are quickly noticed [Craig et al., 1993; de Graaf-Peters and Hadders-Algra, 2006; Giannakoulopoulos et al., 1994; Slater et al., 2006];

3.2 Methods to Assess Fetal Context

Ultrasound machines are devices which emit ultrasound waves to create sonograms, i.e., sound waves above human hearing capabilities (20kHz) are used to create images. Such equipment have different versions with different capabilities each, being available as 2-dimensional, 3-dimensional and 4-dimensional ultrasound (2D-US, 3D-US and 4D-US, respectively) devices. The 2D-US instrument is capable of capturing 2-dimensional images without detailed geometrical volume perspective, while the 3D-US version adds detailed geometrical volume perspective and the 4D-US machine captures videos with geometrical volume perspective on each frame.

Among the advantages of using 4D-US machines to assess fetal context it includes assessing through high-quality images in real-time the fetal movements and facial expressions, and the discovery of prenatal neurodevelopment. Additionally, structural or functional problems may also be evaluated.

Towards the objective of auscultation, or listening to internal sounds in an animal or human body, there are different available devices. Particularly for fetuses, examples include the Pinard horn, which is a type of stethoscope, the Doppler fetal monitor and the cardiotocograph (CTG), which are ultrasound devices, all of them are used to monitor the fetal heartbeat for prenatal care. Fetal health and wellbeing may be damaged by the misuse of ultrasound-based devices through long duration of examinations, poor angulation and position of the sound emitting ends of the system towards the mother's body or emitting sound frequencies outside of adequate range. The misuse could potentially lead to thermal and non-thermal effects on the fetal tissue, including the possibility for over-heating fetal tissue and introducing mechanical stress on the fetus due to cavitation, radiation force, and acoustic streaming [Church and Miller, 2007].

3.3 The Benefits of Early Pain Detection in Human Life

Detection of pain-related behavior in the human fetus during healthy and pathological pregnancies, especially as early as possible, has three main implications:

- One is related to the obvious point that pain-related behaviors and suffering would request not only specific analgesic treatment during intrauterine procedures and surgeries but also a follow-up to assess the efficacy of such analgesic treatments would be mandatory, something that is not currently done;

- Another issue is related to the fact that pain and suffering may occur in certain intrauterine diseases and besides the demand for treatment of the experience of pain itself and the subsequent cardiovascular, metabolic and hormonal changes accompanying it may worsen the health of the fetus and constitute a supplementary burden on fetal wellbeing, in addition to the already existing pathology;

- A third and perhaps the most impacting point is that pain-related behavior is a potential marker of fetal vitality and health. Fetal vitality parameters used in everyday clinics include a group of heterogeneous and highly variable parameters that have as a group low sensitivity and specificity that would welcome strategies to increase its prognostic yield;

Furthemore, pain might be an indicator of health conditions or diseases requiring fetal surgery. There are specific scenarios in which surgical interventions still in the intrauterine phase offer the possibility of improving wellbeing and general life quality for the child after birth.

3.4 The Benefits of Medicine-Based and Surgical Interventions in Fetuses

Many prenatally diagnosed conditions can be treated before birth, ranging from little punctures to major fetal surgery. Such fetal health conditions could be amended, potentially saving the fetus life, improving its intrauterine wellbeing and life after delivery.

Myelomeningocele is an example of fetal abnormality in which the standard intervention method consists of surgery, as presented by the MOMS (Management Of Myelomeningocele Study) trial [Adzick et al., 2011], where the outcomes of pre and after child delivery repair were assessed, leading to the conclusion that the procedure performed in the prenatal phase might result in better neurological function. Later on at school-age (5.9 to 10.3 years old), the children originally present in the MOMS trial were assessed, and despite the fact that there were no significant differences in adaptive behavior, the group that received prenatal repair had significantly better motor function and quality of life [Houtrow et al., 2020].

A study of fetal responses to painful stimuli revealed that painful interventions could have long-term impacts on them [Van de Velde and De Buck, 2012], leading to the conclusion that adequate pain treatment is indicated for potentially painful procedures since it not only improves fetal wellbeing but also supports fetal immobilization, which prevents undesired fetal movements from compromising these procedures.

3.5 The NFCS and its Adjustment for Application to Fetuses

Despite the existence of multiple non verbal pain markers, the NFCS is particularly focused on facial expressions, and was originally proposed as a method for pain evaluation in newborn infants [Grunau et al., 1998]. It may be used to monitor pain in premature and full-term infants, however, it wasn't originally intended to be applied on fetuses. Recent advancements have shown the feasibility of assessing acute pain related facial expressions in the human fetus using 4D-US images [Bernardes et al., 2018] followed by an adaption of the NFCS, the fetal-5 scale, with a cutoff index suggesting the discrimination between painful and nonpainful fetal states [Bernardes et al., 2021].

The idea behind the fetal-5 scale is that facial expressions, in addition to the extra non-facial feature neck deflection, monitored using 4D-US machines are used to assess whether fetuses are experiencing pain. Furthermore, each facial expression is assumed to be a marker where the co-occurrence of multiple such markers increase the likelihood that a fetus is experiencing pain. The authors have suggested a cutoff index, where the co-occurrence of 5 or more pain markers lead to the conclusion that a fetus is, indeed, experiencing pain, however, less than 5 leads to the opposite conclusion. The assessed fetuses were in the third trimester of pregnancy, where it was assumed that fetuses are capable of presenting facial expressions in addition to experiencing pain.

Notwithstanding, in order to develop the fetal-5 scale, raters blinded to whether fetuses belonged to the acute pain or control groups, scored 65 pictures of fetal facial expressions based on the presence of 12 items: brow lowering, eyes squeezed shut, deepening of the nasolabial furrow, open lips, horizontal mouth stretch, vertical mouth stretch, lip purse, taut tongue, tongue protrusion, chin quiver, neck deflection and yearning. Similarly to what is done in the present work, those 65 images comprise a set of images extracted and filtered from a set of 4D-US video recordings from fetuses belonging to the acute pain group, and control at rest and acoustic startle groups. The 12 items were analyzed and filtered, according to the author's criteria for redundancy and usefulness, excluding 5 items that were considered to be of low discrimination capacity, ultimately leading the fetal-5 scale to be comprised of 7 final items: brow lowering, eyes squeezed shut, deepening of the nasolabial furrow, open lips, horizontal mouth stretch, vertical mouth stretch and neck deflection. Therefore, the authors reduced the number of facial items present in the fetal-5 scale, something similar to what has been proposed to the original NFCS for neonates [Peters et al., 2003]. The fetal-5 scale was named in such way due to 5 being the cutoff value discriminating painful from non-painful fetal states.

Chapter 4

Background on Deep Learning

Deep learning, also known as deep structured learning, is a subfield of the broader machine learning field, which in turn, is also a subfield of the broader artificial intelligence area. Deep learning is focused on models and algorithms which are inspired by the structure and function of the (human) brain called ANNs (Artificial Neural Networks) or simply NNs (Neural Networks). Deep learning has become increasingly popular due to extremely impactful advancements on the field that enabled it to address a broad variety of interesting AI problems while achieving impressive results, even beating previous SOTA (State-Of-The-Art) results at times. Machine learning is a field concerned with models and algorithms capable of learning patterns from data samples, notwithstanding, deep learning also has the same goal, however, it is focused on the usage of ANNs.

An ANN is a mathematical algorithm consisting of groups of connected unit(s), generally called (artificial) neuron(s), in a layered fashion, where those artificial neurons and their connections approximately model the biological neurons and their synapses, respectively. ANNs take input data as numbers and such connections between artificial neurons may mathematically transform the input data which, in turn, contributes towards the activation or non-activation of such artificial neurons that can propagate the obtained result forward. The precise nature of such mathematical transformations along each layer over an ANN actually depends on the exact layer type and its purpose, e. g., layers could be of type fully connected, convolutional or attention, however, their specificities will be properly introduced later on.

ANNs layers such as the fully connected contains kernels which consist of learnable parameters, also called trainable parameters, which are discovered during training time, they are used by the layer to transform any data it receives as an input, producing outputs. Such outputs might also be subject to non-trainable transformations such as the ones performed by activation functions. Inside of ANNs, the raw information is received by the first layer, called input layer, and it propagates forward through a sequence of layers where each of them may transform the data in a unique way, until it reaches the last layer, called output layer, which contains the answer produced by the ANN.

4.1 Machine Learning Paradigms

Diving a bit further into the broader field of machine learning, there are many different slopes of problems that require different approaches in order to be solved. Each learning method is suitable for specific scenarios, based on conditions such as the type of data available or the nature of the characteristics required to be learned, among plenty more. There are 3 fundamental paradigms of automated learning:

- Supervised: consists of problems where the target answer is known for each data sample;
  - Classification: consists of problems where the target outputs are of categorical nature, generally they are represented as probability values in the [0, 1] range;
  - Regression: consists of problems where the target outputs are numbers, generally constrained in intervals;
- Unsupervised: consists of problems where the target answer is unknown for each data sample, therefore, the algorithm's goal is to find answers which best fits certain set of criteria;
  - Clustering: consists of grouping elements that are somehow related;
  - Manifold: consists of learning algorithms that change the original dimensionality of data;
- Reinforcement: consists of learning agents capable of taking actions or making decisions which attempt to maximize the cumulative reward during the agent's existence in the environment;

Note that the above expressions and organization of the different learning tasks might differ from researcher to researcher.

4.2 Fully Connected Layer

The most basic building block of ANNs are fully connected layers, which are also known as dense or feedforward layers. It contains two kernels (sets of trainable parameters), that are used to transform the input, into the output using the following mathematical formulation:

$$output = input \cdot k_1 + k_2 \tag{4.1}$$

In the above formula, k_1 is a multiplication kernel, k_2 is the bias kernel, both of them are matrices representing a set of calculations all in one equation. The fully connected layer may also be understood as a function which maps from $\mathcal{R}^m$ to $\mathcal{R}^n$.

4.3 Convolutional Neural Networks

CNNs (Convolutional Neural Networks) refers to ANNs which are composed of one or more convolutional layers, they were originally designed to work with images in a pixel level, however, they are really good at leveraging spatial relationships from the input data's structural distribution. CNNs have become a powerful tool [LeCun et al., 1999] to learn patterns from spatially dispersed data such as images, where CNNs excel at detecting patterns such as edges, corners, color shifts, among others.

A convolution operation may be performed on data of different dimensions, it consists of a linear combination of neighboring features from the input, in order to produce the output at each convolution layer. The linear combination of input features are determined by the convolution kernel, which is a matrix of weights that convolves all the neighbors of a feature together, notwithstanding, the convolution may be applied to all features, where many kernels may be used. Such convolution kernels, also called filters, are composed of trainable parameters that learn how to extract relevant patterns, when available, in each block throughout the input data.

Another important characteristic which makes CNNs powerful is the capability of extracting feature maps, they consist of relevant patterns particular to the training data. Previously to CNNs in the field of computer vision and pattern recognition, there were segmenter algorithms hand-crafted for general pattern recognition in images, such as the Harris detector [Harris and Stephens, 1988], among others. Since they were hand-crafted algorithms built for general purpose applications, the segmentation and feature extraction process often relied on simplifying assumptions about the input data and could rarely exploit the best patterns that could lead to the best results.

Furthermore, CNNs are frequently associated with pooling operations which is a great method to reduce the data dimensionality. After extracting the feature maps, generally into a lower dimensional space, fully connected layers are used to combine the extracted patterns to arrive at the final predictions.

4.4 Residual Neural Networks

Residual Neural Networks, also called ResNets [Kolen and Kremer, 2001], have their architecture designed in such a way where certain layers might have their output forwarded to other layers ahead, skipping one or more layers along the sequence. The residual characteristic of such ANNs lies in the fact that residual output values from previous layers are forwarded, taking shortcuts to reach deeper layers while skipping layers in the natural sequence. Such architecture is often accompanied by batch normalization, a layer that learns how to scale its inputs by adjusting it into a gaussian distribution of mean 0 and variance 1, ultimately equalizing the scales of outputs provided by the concatenation of two or more different layers.

In addition, the deeper neural networks grow the more difficult they become to be trained, therefore, ResNets were developed to overcome such difficulties, mitigating the problems of vanishing gradients [Pascanu et al., 2013] and degradation (accuracy saturation) [Rakitianskaia and Engelbrecht, 2015]. Furthermore, the first problem arises with gradient-based ANN optimization methods, backpropagation, and deep neural network architectures, where the trainable parameters involved receive an update proportional to the PDEF (Partial Derivative of the Error Function) with respect to the current weight in each iteration of training. The PDEF accumulated across several stacked ANN layers, responsible for the learning process, might present small scales and, therefore, it may vanish before the backpropagation process is capable of reaching and updating all layers. This behavior during training time prevents relevant updates that would otherwise contribute to the model learning and improving, ultimately leading to deep learning models that can't be trained any further or at all. Additionally, the second problem arises with degradation and it refers to the state in which a neuron predominantly outputs values near the asymptotic ends of the bounded activation function.

4.5 Multi-Head Attention System in Neural Networks

Given a set of tensors named query, keys, values, and output, the mapping of a query and a set of key-value pairs to an output can be characterized as an attention function [Vaswani et al., 2017]. The result is a weighted sum of the values, with the weight allocated to each value determined by the query's compatibility function with the corresponding key. This particular form of attention uses what the authors call scaled dot-product attention and, given the tensors V, K, Q for value, key and query, and the number of dimensions d_k of K, it is computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4.2}$$

The multi-head attention system is defined as multiple attention heads executing in parallel. The random initialization of the attention heads might lead each of them to learn particular characteristics different from one another.

4.6 Dropout Regularization as a Method to Prevent Overfitting

Deep neural networks containing a huge number of trainable parameters are very powerful and relevant to properly solve interesting problems, however, overfitting is a severe issue limiting their usage, therefore, the employment of regularizers are indispensable in such scenarios. Dropout [Srivastava et al., 2014] is a regularization strategy that prevents complex co-adaptations of the trainable parameters on ANNs over the training data, decreasing the overfitting phenomenon. It's a really effective approach to use in neural networks during training time in order to accomplish more balance in the relevance a model gives to each feature, i.e., given the combinatorially large variety of internal contexts in which neurons must operate, they learn to detect a feature that is generally helpful for producing the correct answer [Hinton et al., 2012].

Dropout is a strategy in which a subset of neurons is neglected during training, being dropped out at random. As a result, on the forward pass, their contribution to the activation of downstream neurons is momentarily removed, and on the backward pass, any weight updates are not applied to them. Many approaches for preventing overfitting have been devised, including terminating training as soon as performance on a validation set begins to deteriorate, applying various weight penalties such as L1 and L2 regularization, among plenty more.

4.7 The Binary Cross-Entropy Loss Function

Machine learning models learn how to map a set of inputs to a set of outputs based on training data over a mathematical optimization process, however, to accomplish the discovery of well suited mapping functions, it is necessary to be able to compute how bad a model is performing, i.e., the error, also called loss. It might seem superfluous and unnecessary to do so in complicated ways, however, there are many scenarios where the method for computing such loss is extremely impactful on the models' final performance.

In this work we focus on a particular loss function called binary cross-entropy, which is focused on classification problems where predictions are to be made between only two possible classes. The method is based on ideas from information theory, which in summary, when an event is less likely to be observed, it includes more information. On the opposite hand, when an event is more likely to arise, it carries less information. Therefore, the entropy of a random variable is the average level of information, i.e., uncertainty, inherently present in the variable's possible outcomes. The cross-entropy function is based on the concept of entropy, from information theory, and it computes how far average events are separated from the distributions of two random variables.

4.8 Training, Validation and Testing

Training is a process through which machine learning models learn over the training data, i.e., the optimization process attempts to minimize the loss function values over each training batch of data sample on each training iteration. Particularly in deep learning, training consists of iteratively adjusting the values of all trainable parameters, ultimately reducing the loss values until reaching the best generalization capabilities.

Additionally, there are limits to the extent to which any machine learning model can learn, since such models represent mathematical functions which attempt to approximate an unknown oracle function. It's perfectly common to have scenarios where the chosen models offer poor flexibility to learn in depth from data of certain problems, ultimately leading to models that slightly outperform random decisions. Further, it's also common to find scenarios where models strongly outperform random decisions, however, they also present a tendency towards stagnation of quality metrics below perfection.

Furthermore, frequently machine learning training behaves as an extremely sensitive procedure which comprises many different learning stages that determine how models will perform on unseen data subsequently. Further, models which train for too few iterations could suffer from underfitting, while antagonistically, training for too many iterations leads to overfitting. Such training issues are generally undesired and harmful to models' performance when attempting to obtain generalization capabilities, except on specific applications or novel machine learning techniques that can particularly benefit.

Moreover, validation is a process that involves separating data, generally taken as a small set of samples from the training data, in order to evaluate how the models' quality metrics would perform on it executing inference. This procedure is advantageous to guide hyperparameters tuning, which also includes architecture definitions on deep learning models.

Nonetheless, testing has a different objective when compared to validation. Testing aspires to produce quality metrics for the trained models as if it was executing inference on novel data that it has never had any contact with previously. In other words, testing is concerned with understanding how trained models would perform, in inference mode, while deployed to production in the real world.

Leaking the same samples between the training, validation and testing sets of data are generally statistically undesirable, since it could lead to untrustworthy quality metrics. On this matter, quality metrics would often present better results than what was actually intrinsically achieved by the trained models. This phenomenon occurs, for instance, because if certain data samples are used for training it means the loss is minimized on such samples, therefore, if the same samples appear in the testing set the model is further than normal inclined to more accurately make predictions.

4.9 The VGG-Face-16 Pre-Trained Model

The VGG-Face-16 (VGGF16) model is a deep learning model with an architecture based on a CNN [Parkhi et al., 2015]. It was trained by the VGG (Visual Geometry Group), an academic group focused on computer vision at Oxford University, on a very large scale dataset containing 2.6M images, and over 2.6K people. The dataset was assembled by a combination of automation and human in the loop, the resulting model achieved comparable state of the art results on the standard Labeled Faces in the Wild [Huang et al., 2008] and the YouTube Faces [Wolf et al., 2011] datasets. The employed keras version of the VGGF16 used in the present work receives as input a batch of images with the default shape of $(224, 224, 3)$ in channels last format, in addition, it possess a standardization pre-processing function which adapts the input images to the same particularities of the data it was originally trained on.

Frequently the fully connected layers existing at the end (also called top) of the VGGFace16 ANN are ignored when using the pre-trained model for different applications. That's because the latent space features provided as the output of the convolutional and pooling blocks are feature vectors carrying higher levels of relevant information about the raw input data, enabling them to be fed into other models, including those non ANN based.

4.10 Transfer Learning and Fine Tuning

Properly training machine learning models enable them to become experienced on the training data, intrinsically empowering them to detect patterns on such data. Notwithstanding, there are groups of problems which have similar characteristics, for example, predicting whether it's going to rain today or the volume of rain water to be expected are, indeed, different but still very similar problems from a machine learning based perspective. Therefore, transfer learning is a technique which consists of training models to solve a problem A, then using fully or partially those trained models as a starting point to train them once again on another similar problem B. The goal is to exploit patterns learned by machine learning models on problem A in order to facilitate the learning process on the later problem B, ultimately leading to a better performance on problem B when compared to a baseline where models aren't trained on problem A to begin with.

There are scenarios where using pre-trained models for another similar problem promotes great advantages, for example, if problem A possesses plenty of data samples available, however, problem B possesses too few samples. Accordingly, it would be feasible to directly solve problem A while the same isn't true for problem B due to the amount of training data samples available for each problem, therefore, the transfer learning technique could be employed in case problems A and B are similar enough. Determining beforehand whether two problems are similar enough for the transfer learning technique isn't trivial in many cases, thus, empirical experimentation may be employed. Furthermore, the process of training models on problem B after they were trained on problem A is called fine tuning. The fine tuning process on problem B is commonly performed slower than the initial training procedure on problem A, since it could reach the overfitting state with fewer than normal training iterations.

4.11 Data Augmentation

Data augmentation (DA) is a term that refers to various strategies for increasing the size and quality of training datasets so that stronger machine learning models may be trained using them [Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019]. The data subject to augmentation techniques may be presented in numerous forms such as tabular, signals, images, videos, among plenty more. Each particular data format and machine learning application enables different augmentation techniques.

Moreover, this dissertation is focused on augmenting images by applying transformations that change properties from them such as rotations, translations, shearing, color intensity shifts, brightness shifts and horizontal flipping. Such images have their nature altered in a constrained manner, while having their associated label remaining the same. Figure 5.4 presents an example of DA which was actually employed on this project.

4.12 Explainability

Explainable Artificial Intelligence (XAI) is a field concerned with enabling humans to understand the output provided by artificial intelligence systems. The concept of *black box* refers to artificial intelligence systems that doesn't ballast its solutions with the possibility of providing explanations to its users and interested parties, withholding reasoning details regarding how the output solution was achieved. Despite the differences in reasoning between machines and humans, explainable AI systems are generally preferred over *black box* ones, because adequate explanations are critical to improve how trustworthy AI systems should be assumed to be, since they are prone to complications caused by a series of conditions such as outliers due to data impurity, underfitting, overfitting, among other forms of biases which ultimately leads to poor quality results. Through explainable AI it is possible to validate and challenge existing human knowledge, in addition to create novel assumptions, something that fuels the virtuous learning cycle between humans and machines.

4.13 Quality Metrics

Quality metrics, also called performance metrics, refers to defining quantitatively how much machine learning models can achieve on unseen data samples, i.e., data samples which weren't used for training. It's extremely important to understand when models are improving in addition to being able to compare them.

Moreover, multiple metrics exist and some of them are actually used as loss functions, while others can only be employed for evaluation, as opposed to being part of the optimization process. On that matter, any metric being used as a loss function for deep learning is required to be differentiable, so the gradient descent process may consistently take place, while it isn't mandatory for metrics intended for pure evaluation to fulfill this requirement.

Furthermore, there are plenty of available metrics such as ROC AUC (Receiver Operating Characteristic Area Under the Curve), binary accuracy, cosine similarity, mean squared error, among more. Each of them might have restrictions on the nature of problems they can be used to evaluate on, with their own scales and quality direction, i.e., they might become better as they decrease in some cases or increase in other cases.

Chapter 5

Automatic Recognition of Painful Facial Expressions in Fetuses

This chapter presents the first component of this project's methodology, comprehending data collection, labeling, in addition to multiple automatic and human-in-the-loop preprocessing steps.

5.1 The Acquisition of Medical Data

As shown by the Fetal Pain Study Group (FPSG) from Universidade de São Paulo (USP), it is feasible to record fetal facial expressions using a 4D-US machine [Bernardes et al., 2018], consecutively, they have recorded 13 fetuses each in one out of three conditions [Bernardes et al., 2021] in a room set-up according to Figure 5.1:

1. Acute Pain (AP) group: fetuses with diaphragmatic hernia (fetoscopic endoluminal tracheal occlusion) with indication to surgery still in utero were evaluated during the anesthetic injection into the thigh in the preoperative period;

2. Control at Rest (Co-Re) group: during shceduled 4D-US examinations, after a 5-minute rest time for the mother, resting fetuses were recorded in a calm and dark room;

3. Control Acoustic Startle (Co-AS) group: Fetuses were recorded while exposed to acoustic stimuli, where such distresses were utilized to enhance the accuracy of fetal heart rate monitoring, which was used to determine fetal well-being. The stimulator used was comparable to a bicycle horn, and it was placed to the maternal abdomen for 4 seconds adjacent to the fetal cephalic pole, producing 3

pulses of acoustic waves between 500 to 4000 hertz with intensity between 60 to 115 decibels.

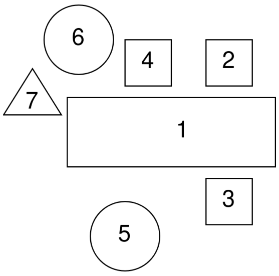


Figure 5.1. Set-up for surgery and face recording in an operating room. (1) The mother's position; (2) the chief surgeon who performed the puncture; (3) the assistant surgeon who obtained the 4-D images; (4) the surgical technologist; (5) the ultrasound machine used in surgery to focus the fetal trachea/thigh; (6) the ultrasound machine used for fetal face recording; and (7) an external camera [Bernardes et al., 2018]

Fetuses in the AP group were recorded previously and subsequently to the anesthetic injection, similarly, the Co-AS group was also recorded before and after the acoustic stimuli. The AP, Co-Re and Co-AS groups were comprised of 5, 4 and 4 fetuses, respectively. All assessed fetuses were in the third gestational trimester i.e., 31.1 ± 2.8 weeks into the pregnancy, and mothers were 28.7 ± 5.5 years old. Further, there were no concurrent neurological abnormalities or illnesses in the AP group, in addition, neonatal checkup after delivery indicated that the fetuses without congenital illness belonging to the Co-Re-AS groups were indeed healthy.

5.2 Extracting Images from Videos

The original data used in this work is the same collected by the FPSG and it was presented in the format of 13 4D-US videos, each of them contextualizing a single fetus. Further, in favor of developing the new machine learning application, it was necessary to formulate a couple pre-processing steps. Therefore, we sampled from the sequence of images constituting the videos, however, for each video it's corresponding particular sampling rate was employed. Despite all videos having 30 frames per second, the sampling rate for each particular video was manually set to attempt promoting balance between the total amount of samples, the amount of samples for each of the pain and non-pain classes and quality of the images.

5.3 Segmentation of a Binary Classification Problem: Labeling Data

Differently from the groups created by Bernardes et al. [2021], the present work created different groups focusing on building a novel machine learning dataset comprised of 4D-US images of fetal facial expressions. Fetuses were divided into 2 groups, Pain and Non-Pain, being the first one comprised of painful and the second one of painless fetal images.

Based on the original 4D-US videos collected by Bernardes et al. [2021], the AP group was subject to an acute pin prick, while the Co-AS group was subject to acoustic stimuli. These events are important markers that have been used in the present work to form the novel Pain and Non-Pain groups according to the configuration presented by Table 5.1 and Figure 5.2. There were no events of interest for the Co-Re group.

Table 5.1. Original groups of fetuses [Bernardes et al., 2021] sourcing images to the novel Pain and Non-Pain groups. The values true and false determine whether images from the group of fetuses identified in the column were inserted into the new pain or non-pain groups identified in the rows.

New Groups	AP pre-stimulus	AP post-stimulus	Co-Re	Co-AS pre-stimuli	Co-AS post-stimuli
Pain	false	true	false	false	false
Non-Pain	true	false	true	true	true

Since the acute pin prick stimuli received by the AP group possesses anesthetic effects, an upper limit of 10 seconds post-stimuli was set, ultimately preventing unreliable images of anesthetized fetuses to be added to the new Pain group. The anesthetic effects on third trimester fetuses could bias any indicators of pain, therefore, narrowing the time window used to collect painful facial expressions after the painful pin prick event mitigates this problem. All images beyond the tolerance time window were discarded, since the resting state induced by the anesthesia could bias any of the new Pain and Non-Pain groups in case they were to receive such unreliable images.

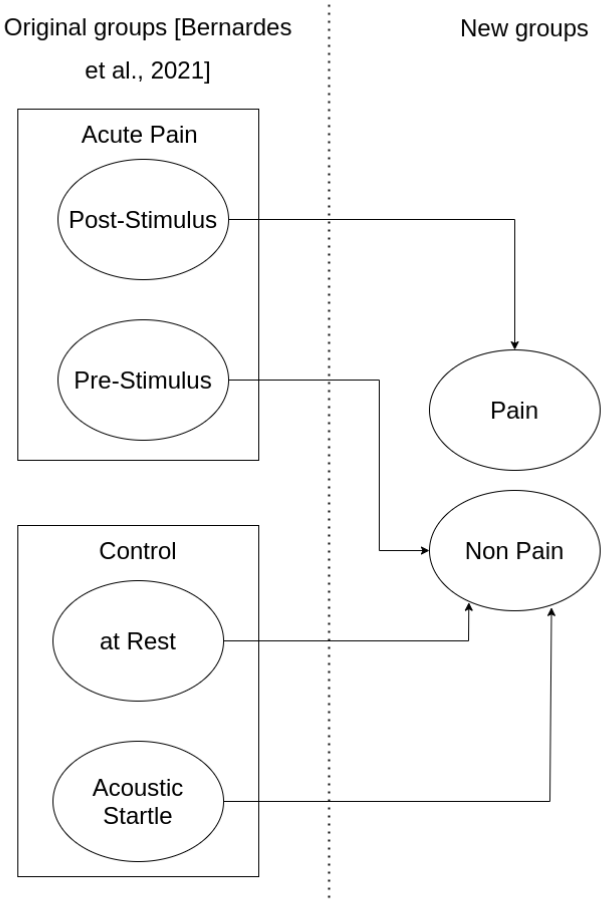


Figure 5.2. Labeling of the original image samples. The original groups of fetuses [Bernardes et al., 2021] identified on the left of the dotted line were inserted into the new groups, presented on the right hand side.

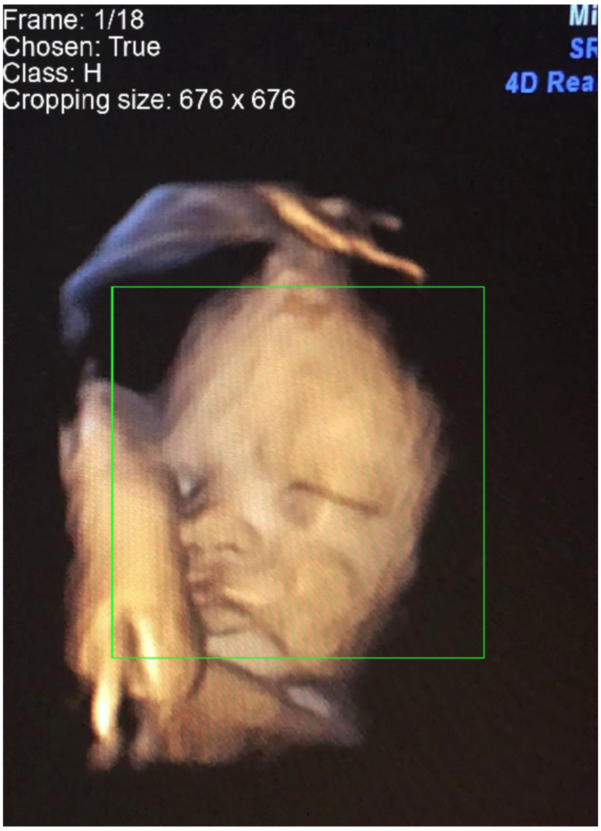
Moreover, the events of interest for both the AP and Co-AS groups weren't instantaneous, they took place along a short time window. The images classified as pre-stimulus are originated exclusively before the instant where the stimulus start, by another hand, the images considered post-stimulus took place inclusively afterwards. Therefore, post-stimulus images might overlap in time with the actual duration of the events of interest and go beyond it.

5.4 Frame Cropping and Assisted Manual Filtering

Ideally, the created datasets should be large enough to enable the computer vision machine learning model to achieve its fullest potential, however, that's a challenge due to the lack of large data samples. In addition, frames constituting the original raw videos were sampled in the scale of many per second, specially for the painful scenario, ultimately leading to many similar samples, since they are temporally related. Therefore, cropping the sample images to focus on the fetal faces and filtering out poor quality images were extremely important steps, in addition to the techniques discussed in Section 5.5.

The main purposes to crop images are to (i) reduce the amount of potential artifacts unrelated to facial expressions present in the images, ultimately constraining the focus of the vision models, and (ii) promote greater similarities between the conditions in which faces were presentend in the datasets created in the present work and the one used to train the VGGF16 feature extraction component. These efforts have the potential to increase the quality of the latent features extracted by the VGGF16 component.

Notwithstanding, the process of recording fetal facial expressions was subject to issues arisen by movements of the fetus (1) during the ultrasound session, movements of the probe of the 4D-US device (6) performed by the ultrasonographist, and movements of the external recording device (7), according to the items listed by Figure 5.1. Therefore, the original raw videos provided samples where the fetus' face wasn't properly visible, and thus, according to the criteria of a human judge, many samples were discarded. The main criteria utilized to filter out samples was whether the human judge could identify the fetal face along with its elements such as mouth, nose and eyes, without using the temporal aspect of the videos to identify such details, since the ML application won't have the temporal information either. The idea behind ensuring that enough facial elements are present is to avoid using samples that doesn't contain the target facial expressions information, since it is desired to promote an analysis of the explanations generated by the ML models in addition to crossing them with the facial items from the fetal-5 scale defined by Bernardes et al.. Furthermore, throughout the videos the fetus' face switch between visible and poorly visible or even totally invisible, however, the temporal aspect of the videos empower the human judge to track where the fetal face is likely to be during a transition between the visible and invisible condition, something that could jeopardize the performance of the ML application since it wouldn't benefit from the temporal information. On this matter, it's important to highlight that images in which the fetal faces are visible to the human judge are certainly feasible to be detected by the computer vision models as well, however, on the opposite hand, cases where the fetal faces were invisible to the human judge were assumed to be pose a threat to the ML application since it's unknown whether they are feasible to be detected by the ML models.



Frame: 1/18
Chosen: True
Class: H
Cropping size: 676 x 676

Figure 5.3. Cropping and filtering assistant software tool, it was developed particularly for this project. The writings in white on the top left are presented by the tool, indicating the current and total amount of frames in the set being evaluated, whether it was chosen, the original class it belongs to and the current size in pixels of the green selection box surrounding the fetus' face. The green selection box is set by a human user of the tool and it may be done multiple times, further, the software automatically sets such green box to be a perfect square, until one of them is considered adequate to actually crop the image. Once cropped, the image is sent into another directory along with all the other selected samples. Note that the blue text on the top right is considered noise for this project, generated by the 4D-US system.

Moreover, the assisted nature of the manual filtering of images is due to the fact that this work contemplates a software to assist the human judge to view fetal images, frame the fetal face with a perfect square window, crop and save it, according to Figure 5.3. The tool also includes an easy method to quickly decide whether an image being analysed is to be discarded or saved. The images selected to move onto the next phase are resized to match the VGGF16 component input size, i.e. $(224, 224, 3)$ in channels last format and pre-processed to its image standard values. Finally, the image samples used by the present work are not necessarily the same utilized by Bernardes et al., since the entire pipeline was built in an independent manner and meant to contain plenty more resources than just the 65 pictures they have utilized.

5.5 Transfer Learning and Data Augmentation

On the point of training ML models for medical applications, typically, obtaining data in large volumes is a challenging problem. Furthermore, detecting complex facial patterns on fetuses is a difficult problem, which could be mitigated and potentially solved by increasing the volume of training data, however, it isn't the scenario the fetal pain project is up against. The lack of data could become a major problem, however, there are alternatives to mitigate it: TL (Transfer Learning) and DA (Data Augmentation).

In ML, the technique of TL is based on training models for specific problems, then using such models as a starting point to train it once again on another, however, similar problem. Since the assessed fetuses in this project were aged enough (in the third trimester of gestation), they had already developed plenty of facial similarities when comparing with human adults. Therefore, we chose a pre-trained ML model, the VGGFace16, as a starting point for this project to mitigate the small amount of training data available. The VGGF16 model has proven to work surprisingly well for recognizing facial expressions in 4D-US images of fetuses, even though it was originally trained to recognize adult's faces on traditional pictures, i.e. taken by traditional cameras. It's important to note that images captured by 4D-US machines and traditional cameras are very different regarding their nature and the environment where they are employed. Accordingly, the training process consisted on the adaptation of our custom discriminant model to the outputs the VGGF16 could provide by being trained in a frozen fashion.

Further, the technique of data augmentation consists of artificially creating new data samples based on the original input training data. Ideally, such novel samples don’t change the intrinsic meaning of the original samples on which they were based, while adding more diversity. This ultimately enables the trained models to become more robust at identifying the desired patterns. In this project, the meaning of the original images which we are interested in preserving is the presence of the face of a fetus along with all the associated facial expressions. Therefore, augmentations on those images such as changing their brightness, color intensities and rotations were designed preserve the target information while creating data diversity which could improve the quality of our deep learning models.

Table 5.2. Data augmentation hyperparameters: (Level) given augmentation names; (Rotation[1]) in degrees; (Translation[1]) scale relative to the frame's width and height; (Shear[1]) in degrees; (Channel shift intensity[1]) absolute RGB pixel value ranged in the fully closed interval $[0, 255]$; (Brightness[1]) relative to the current brightness; (Flipping) true when a random flip was allowed with 50% of chance, false otherwise, horizontal and vertical, respectively. [1]: Randomly ranged hyperparameters in a fully closed interval.

Level	Rotation[1]	Translation[1]	Shear[1]	Channel shift intensity[1]	Brightness[1]	Flipping
Easy	[-30.0, 30.0]	[0.95, 1.05]	[-5.0, 5.0]	[-12.8, 12.8]	[0.975, 1.025]	t / f
Light	[-30.0, 30.0]	[0.95, 1.05]	[-5.0, 5.0]	[-19.2, 19.2]	[0.950, 1.050]	t / f
Medium	[-30.0, 30.0]	[0.95, 1.05]	[-5.0, 5.0]	[-25.6, 25.6]	[0.900, 1.100]	t / f
Strong	[-60.0, 60.0]	[0.90, 1.10]	[-5.0, 5.0]	[-32.0, 32.0]	[0.850, 1.150]	t / f

The datasets were generated based on randomly chosen hyperparameters, where the hyperparameters were constrained according to Table 5.2. The randomness factor wasn't built in a reproducible way, instead, the generated datasets were saved for posterior usage. Figure 5.4 presents samples collected from the training sets across folds of the augmentation level medium datasets.

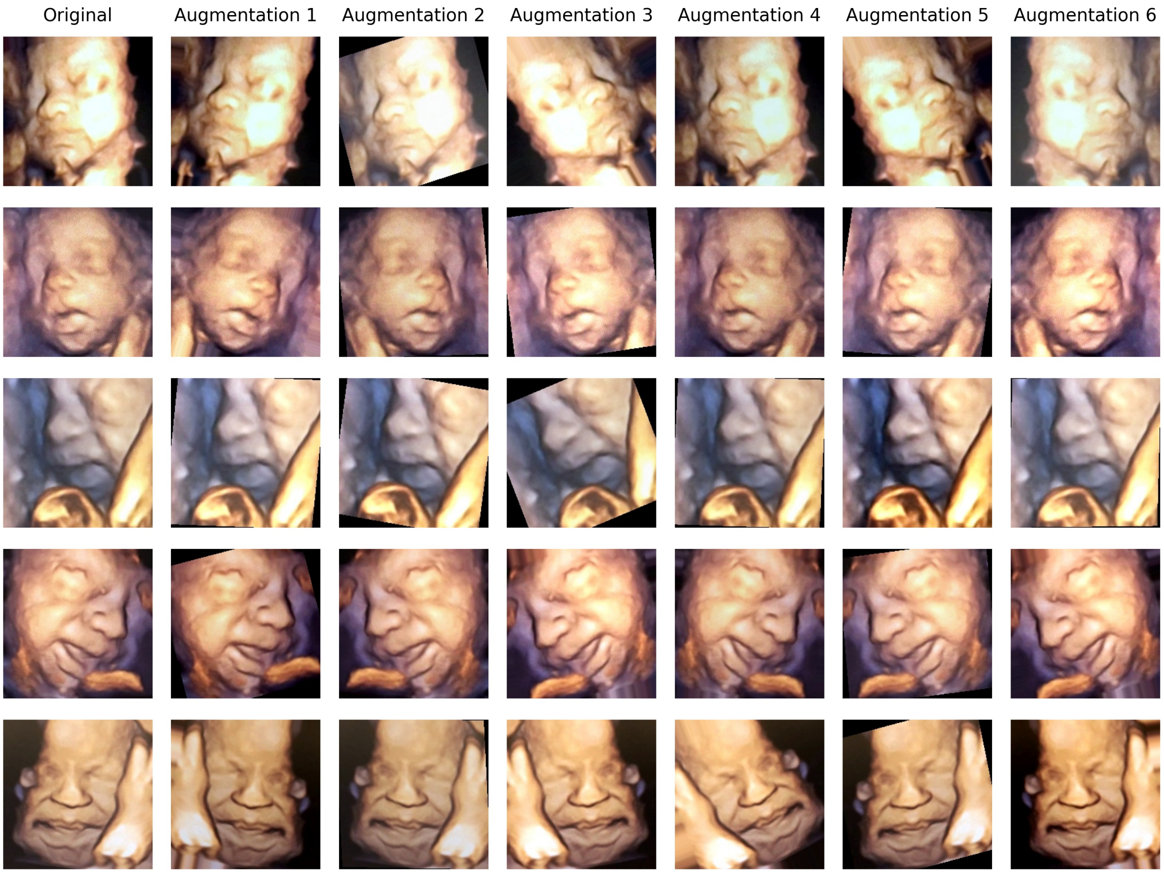


Figure 5.4. Samples of data augmentation applied on fetal images. Each row contains a single fetus, where the leftmost column presents original images and the following columns presents transformations applied on them. All the 5 fetuses displayed belong to the Pain group and were collected from the training sets across multiple folds of the medium augmentation dataset. The complete training sets of each fold might contain up to an original image in addition to up to 10 augmented versions of itself.

5.6 Cross Validation Ready Datasets

The produced datasets have been prepared ready to perform k-fold cross validation, with training and testing sets, however, a validation set wasn't created due to the lack of data samples. The k value for the k-fold is 5, since we only had 5 fetuses recorded in both resting and painful scenarios, therefore, each of them was chosen to figure in the test group once, while the all fetuses belonging to the Co-Re-AS groups were always used only for training. That's due to the small volume of data available, we had to increase as much as possible the number of data samples in the training set. Consequently, since it was necessary to have painful samples for testing, all the samples collected from the same fetus previous and posterior to the acute pin prick had to be placed in the testing group, being a necessary measure to avoid information leakage [Kaufman et al., 2012].

Information leakage consists in the use of information in the model training process that would not be anticipated to be accessible at prediction time, leading the testing predictive scores to overestimate the model's efficacy [Kaufman et al., 2012]. Accordingly, we assumed that similarly to a grown human adult, the fetuses in this study already had developed enough facial characteristics to be uniquely distinct from other human individuals be them either fetuses, children or grown adults. Therefore, we employed the technique of leave-one-out, traditionally used for small datasets, where the left-out element refers to each fetus, each on their respective fold, thus, all images of the same fetus were separated for each testing set. The absence of validation groups or multiple fetuses in the testing groups are a consequence of the low amount of total data available, otherwise the training group would be left with too few samples, ultimately leading to the tradeoff of potentially improving the tests' correspondence with reality at the expense of potentially decreasing the model's actual performance.

Furthermore, the datasets, across all folds and training and testing groups, were produced respecting a perfect class equilibrium of 50/50% between samples labeled as painful and non-painful. Such aforementioned characteristic is beneficial for training deep learning models, since it stimulates the model to update it's internal numerical weights in a class balanced manner, favouring more balanced and robust inferences later on at production time. The total amount of Pain and Non-Pain samples extracted from each video varies, since the class equilibrium used for training was considered globally among all videos, to form the final training and testing groups. On the opposite hand, when each fetus from the AP group entered the testing group the class equilibrium was achieved by eliminating samples from the class with the higher volume of samples until balance was achieved, implicating that high accuracy scores indeed reflect that the resulting models are performing well. For a uniformly random decision model the expected accuracy would be 50%, therefore, we aim our ML models at achieving accuracy scores well above this baseline.

The testing samples for all datasets and their folds weren't augmented, since this process could bias any testing results computed using such samples. The data augmentation was applied exclusively to the training groups as means to improve the generalization capabilities learned by the trained models, however, if the testing set contained the same augmentations, it could become easier for the models to output the correct answer for each image sample. Further, the non-augmented images are more similar to the ones that would potentially be used in a real world application of the present work's ML models and methodology, except for the manual image cropping mentioned in Section 5.4.

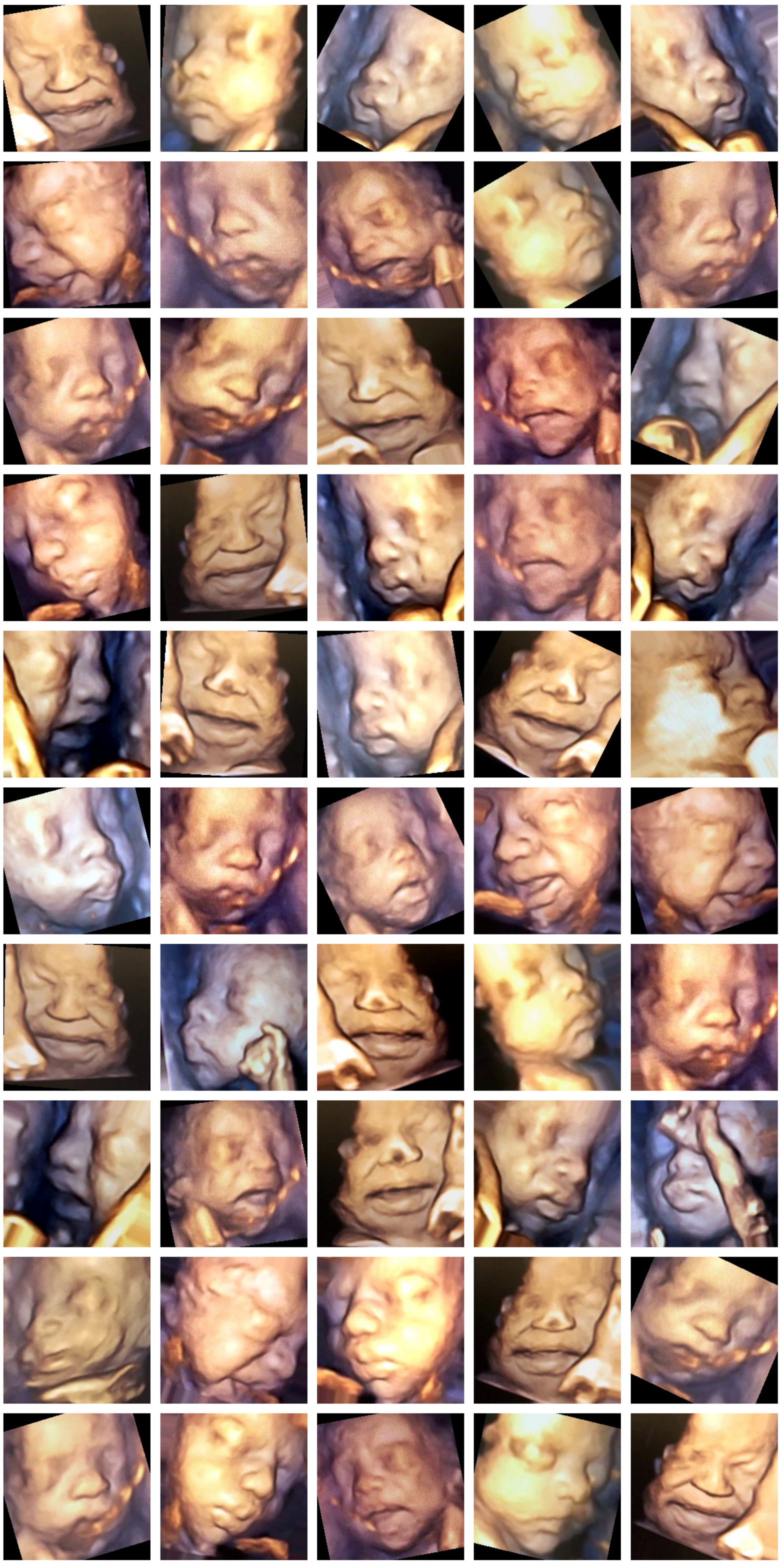


Figure 5.5. Presentation of 40 samples randomly drafted from the training sets of the medium augmentation level to illustrate the data augmentation results.

5.7 Explainability Results and Discussion

In association with the deep neural network architecture presented in Section 6.1, we propose the adoption of a few algorithms to conceive explanations. The employment of multiple explanability methods should mitigate issues each of them carry, such as being hypersensitive to changes in hyper-parameters. Since we are working with images, heatmaps will be used as a visualization method to assess the relevance of areas across each image to the model's output.

In the context of visual processing, saliency in images refers to distinctive properties such as pixels or image resolution, where these unique qualities describe the visually appealing regions in an image, being the saliency maps topographical representations of them. The method of saliency maps proposed by Simonyan et al. [2014] measures the spatial support of a particular class in each image, providing a method for investigating hidden layers in CNNs that can be understood. The map is created by superimposing outcome gradients over the input image, indicating the parts of the photos that were crucial to the classification.

Furthermore, SmoothGrad introduced by Smilkov et al. [2017] is a simple method that can help visually sharpen gradient-based sensitivity maps. It computes the gradients with respect to the input image multiple times, for each of them, noise gets added and the average of the gradients of a large number of inputs that are quite similar to original is taken, easing the inconsistency of the gradient function, leaving only the overall tendency of the gradients across the picture's area. This is assumed to be the most important XAI algorithm for this work due to its intrinsic nature of eliminating noise and capturing the overall gradient trend.

Additionally, there is the Gradient-weighted Class Activation Mapping (Grad-CAM) [Selvaraju et al., 2020], it uses gradients flowing into the final convolutional layer to create a coarse-grained localization map emphasizing key regions in the image that contributes towards the model's prediction.

Chapter 6

Experimental Methodology and Results

This chapter presents the final component of this project's methodology, based on the initial phase presented by Chapter 5. The target application is described in details, along with the presentation of the obtained results.

6.1 The Model Architecture

The model architecture and topology is presented as a diagram in Figure 6.1, it is based on the very deep convolutional block from the VGGFace16 model, without the head (also called top) component. It takes as input batches of images in the channels last format with dimensions $(224, 224, 3)$, in addition, the input images must be normalized according to the VGGF16 standard, otherwise, predictions could become biased. The maximum pooling 4 and 5 layers are used as outputs from the convolutional block according to Figure 6.1 and they are the 15th and 19th layers, respectively, according to Figure 6.2. The amount of trainable parameters in the head component is roughly 5 million, while approximately 14.7 million were pre-trained and inherited from the VGGF16, resulting in approximately 19.7 million in total. The model architecture is the same for all folds of the cross validation and it was developed using the tensorflow python framework.

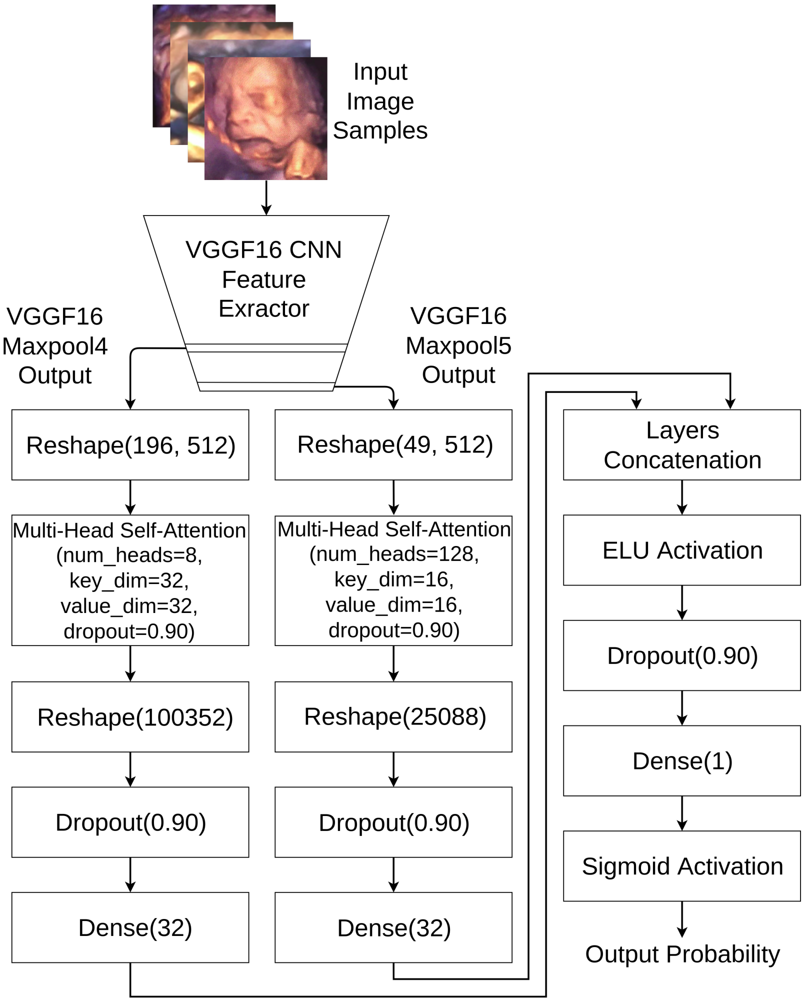
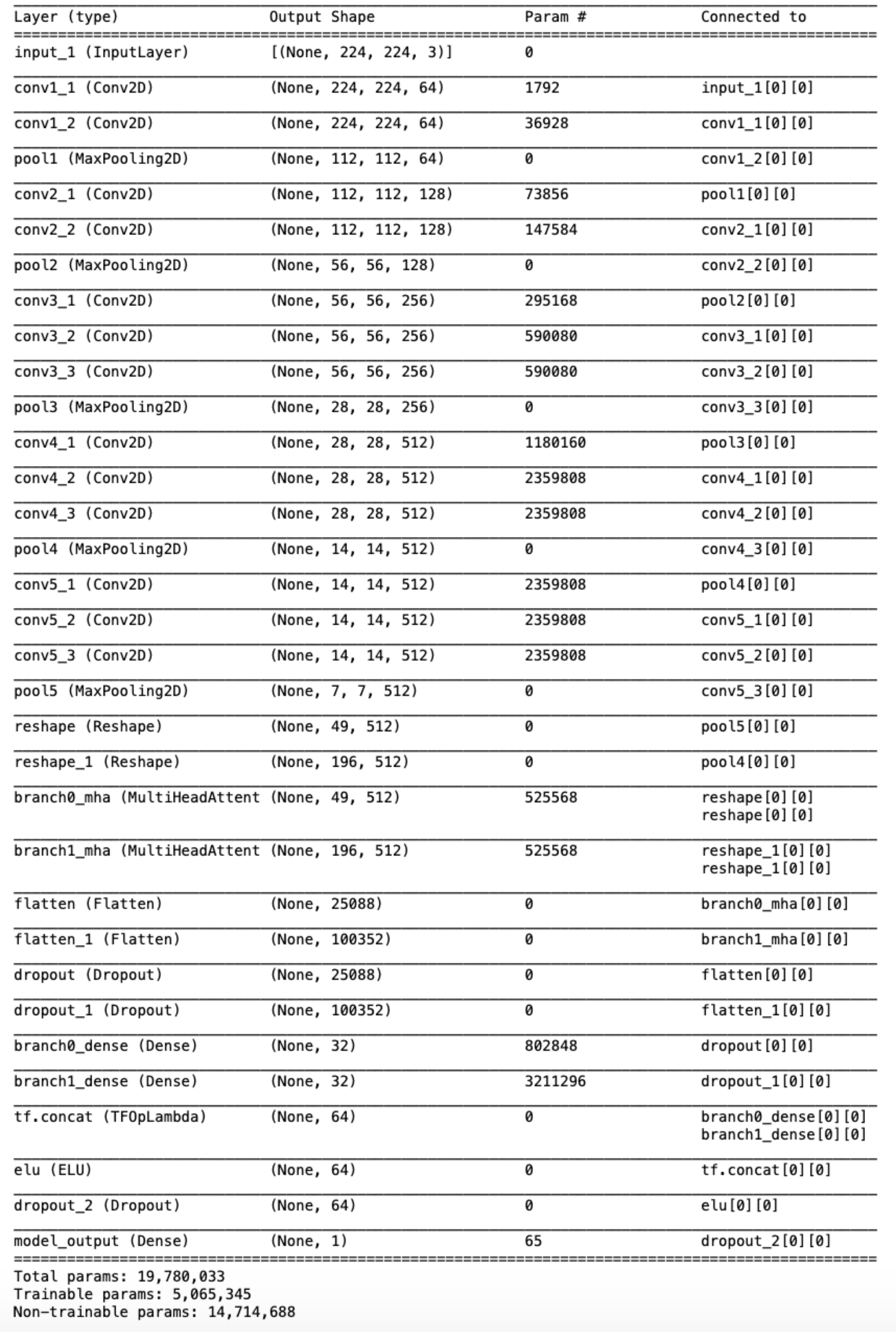


Figure 6.1. The model architecture is based on the VGGF16 CNN feature extraction component, which takes as input 4D-US images, each containing a single fetus, then produces as output features in the latent space. Such features are used by our custom model to output the final probabilities regarding the presence or absence of painful fetal states, the output scale is in the fully open interval $(0, 1)$. The decision threshold for the output probability is 0.5, therefore, the closer the output is to 0.0 or to 1.0 the more confidence in the result for classes Non-Pain and Pain, respectively. On the opposite hand, the closer the output is to 0.5 the less confidence there will be. Dense refers to fully connected layers, while ELU represents the Exponential Linear Unity activation function [Clevert et al., 2016]. The batch size was set to 16 and the learning rate to 10^{-6}.



```
____________________________________________________________________________________________
Layer (type)                   Output Shape              Param #     Connected to
============================================================================================
input_1 (InputLayer)           [(None, 224, 224, 3)]     0
____________________________________________________________________________________________
conv1_1 (Conv2D)               (None, 224, 224, 64)      1792        input_1[0][0]
____________________________________________________________________________________________
conv1_2 (Conv2D)               (None, 224, 224, 64)      36928       conv1_1[0][0]
____________________________________________________________________________________________
pool1 (MaxPooling2D)           (None, 112, 112, 64)      0           conv1_2[0][0]
____________________________________________________________________________________________
conv2_1 (Conv2D)               (None, 112, 112, 128)     73856       pool1[0][0]
____________________________________________________________________________________________
conv2_2 (Conv2D)               (None, 112, 112, 128)     147584      conv2_1[0][0]
____________________________________________________________________________________________
pool2 (MaxPooling2D)           (None, 56, 56, 128)       0           conv2_2[0][0]
____________________________________________________________________________________________
conv3_1 (Conv2D)               (None, 56, 56, 256)       295168      pool2[0][0]
____________________________________________________________________________________________
conv3_2 (Conv2D)               (None, 56, 56, 256)       590080      conv3_1[0][0]
____________________________________________________________________________________________
conv3_3 (Conv2D)               (None, 56, 56, 256)       590080      conv3_2[0][0]
____________________________________________________________________________________________
pool3 (MaxPooling2D)           (None, 28, 28, 256)       0           conv3_3[0][0]
____________________________________________________________________________________________
conv4_1 (Conv2D)               (None, 28, 28, 512)       1180160     pool3[0][0]
____________________________________________________________________________________________
conv4_2 (Conv2D)               (None, 28, 28, 512)       2359808     conv4_1[0][0]
____________________________________________________________________________________________
conv4_3 (Conv2D)               (None, 28, 28, 512)       2359808     conv4_2[0][0]
____________________________________________________________________________________________
pool4 (MaxPooling2D)           (None, 14, 14, 512)       0           conv4_3[0][0]
____________________________________________________________________________________________
conv5_1 (Conv2D)               (None, 14, 14, 512)       2359808     pool4[0][0]
____________________________________________________________________________________________
conv5_2 (Conv2D)               (None, 14, 14, 512)       2359808     conv5_1[0][0]
____________________________________________________________________________________________
conv5_3 (Conv2D)               (None, 14, 14, 512)       2359808     conv5_2[0][0]
____________________________________________________________________________________________
pool5 (MaxPooling2D)           (None, 7, 7, 512)         0           conv5_3[0][0]
____________________________________________________________________________________________
reshape (Reshape)              (None, 49, 512)           0           pool5[0][0]
____________________________________________________________________________________________
reshape_1 (Reshape)            (None, 196, 512)          0           pool4[0][0]
____________________________________________________________________________________________
branch0_mha (MultiHeadAttent   (None, 49, 512)           525568      reshape[0][0]
                                                                     reshape[0][0]
____________________________________________________________________________________________
branch1_mha (MultiHeadAttent   (None, 196, 512)          525568      reshape_1[0][0]
                                                                     reshape_1[0][0]
____________________________________________________________________________________________
flatten (Flatten)              (None, 25088)             0           branch0_mha[0][0]
____________________________________________________________________________________________
flatten_1 (Flatten)            (None, 100352)            0           branch1_mha[0][0]
____________________________________________________________________________________________
dropout (Dropout)              (None, 25088)             0           flatten[0][0]
____________________________________________________________________________________________
dropout_1 (Dropout)            (None, 100352)            0           flatten_1[0][0]
____________________________________________________________________________________________
branch0_dense (Dense)          (None, 32)                802848      dropout[0][0]
____________________________________________________________________________________________
branch1_dense (Dense)          (None, 32)                3211296     dropout_1[0][0]
____________________________________________________________________________________________
tf.concat (TFOpLambda)         (None, 64)                0           branch0_dense[0][0]
                                                                     branch1_dense[0][0]
____________________________________________________________________________________________
elu (ELU)                      (None, 64)                0           tf.concat[0][0]
____________________________________________________________________________________________
dropout_2 (Dropout)            (None, 64)                0           elu[0][0]
____________________________________________________________________________________________
model_output (Dense)           (None, 1)                 65          dropout_2[0][0]
============================================================================================
Total params: 19,780,033
Trainable params: 5,065,345
Non-trainable params: 14,714,688
```

Figure 6.2. Detailed description of the model architecture as a tensorflow keras python object. A detailed description of the VGGF16 convolution block is also presented.

6.2 The Training Process and the Obtained Performance

The deep learning models in this project were trained with the support of an extremely intense regularizer technique, the dropout, across multiple layers. It played an important role in constraining the rhythm at which the model's trainable parameters adapt to the input dataset, easing the search for a better generalization state which might not be achieved otherwise. The deep learning models could easily learn patterns from the training dataset capable of overfitting at the very beginning of the training process. Therefore, a great method for slowing down the pace at which the model learns that doesn't involve abruptly changing the learning rate is using the dropout regularizer across different layers, since the learning rate is tied to the batch size [Keskar et al., 2017].

Optimizing deep learning models on small datasets representing complex problems such as detecting pain based on fetuses' facial expressions require great cautiousness, therefore, another extremely important technique employed was to carefully pick properly matching values for the batch size and the learning rate, where small batch sizes should be matched with small learning rates in order to take advantage of the implicit regularization occurring during the optimization process, since large batches commonly lead to sharp minima, which frequently doesn't generalize as well as flat minima [Keskar et al., 2017][Smith et al., 2021].

Furthermore, each dataset created consists of an augmentation level and is comprised of 5 folds ready for cross validation. Therefore, the produced quality scores for each dataset were obtained by averaging the results across all folds, also enabling the computation of metrics such as standard deviation, variance and confidence intervals. The training history is summarized by Figures 6.3, 6.4, 6.5 and 6.6. In all these charts, the abscissa axis represent the training epochs and the ordinate an absolute score value. Finally, the performance scores were summarized in Tables 6.1, 6.2, 6.3 and 6.4.

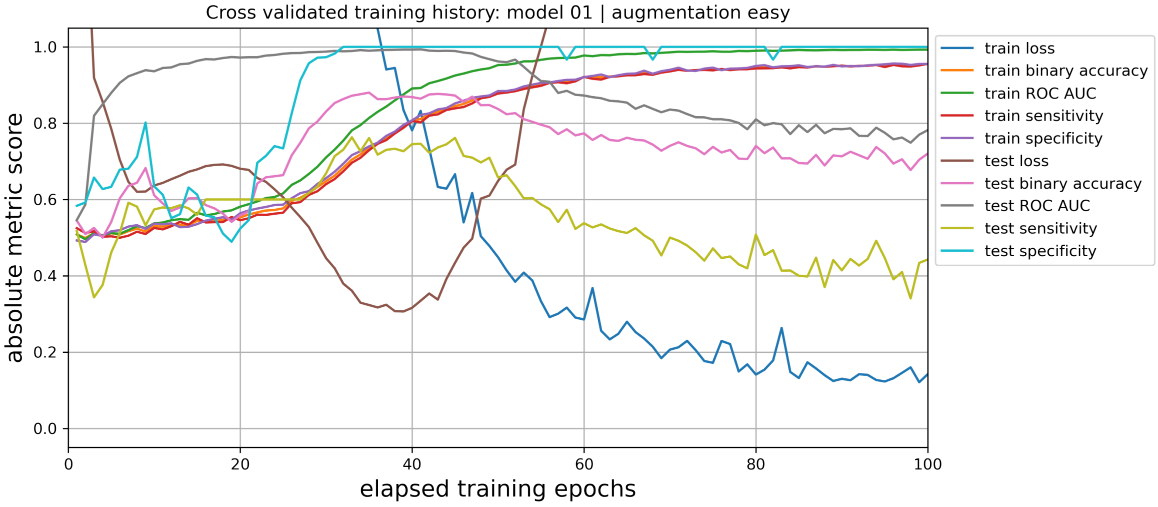


Figure 6.3. Quality scores history for dataset with augmentation easy.

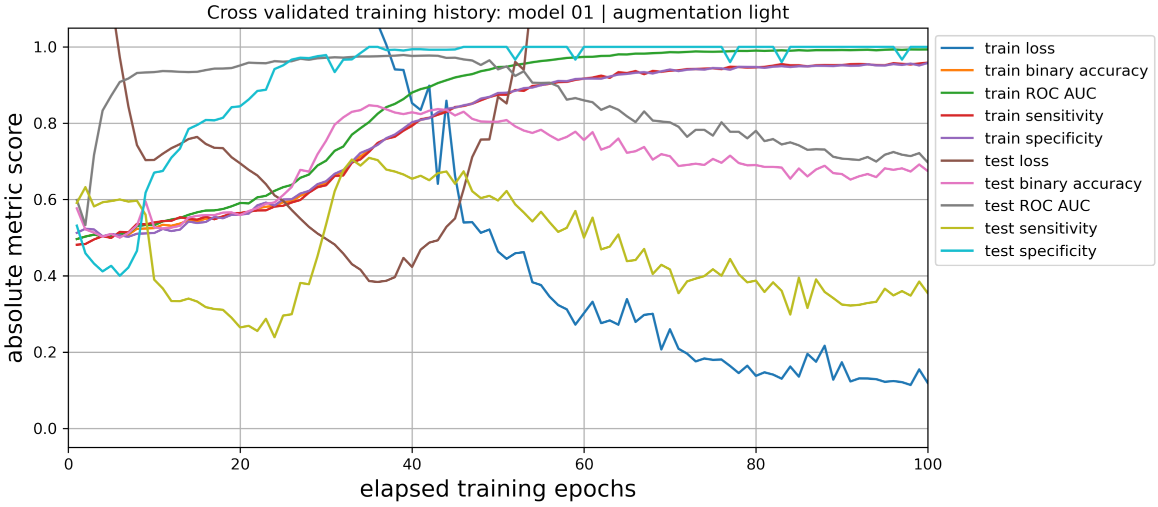


Figure 6.4. Quality scores history for dataset with augmentation light.

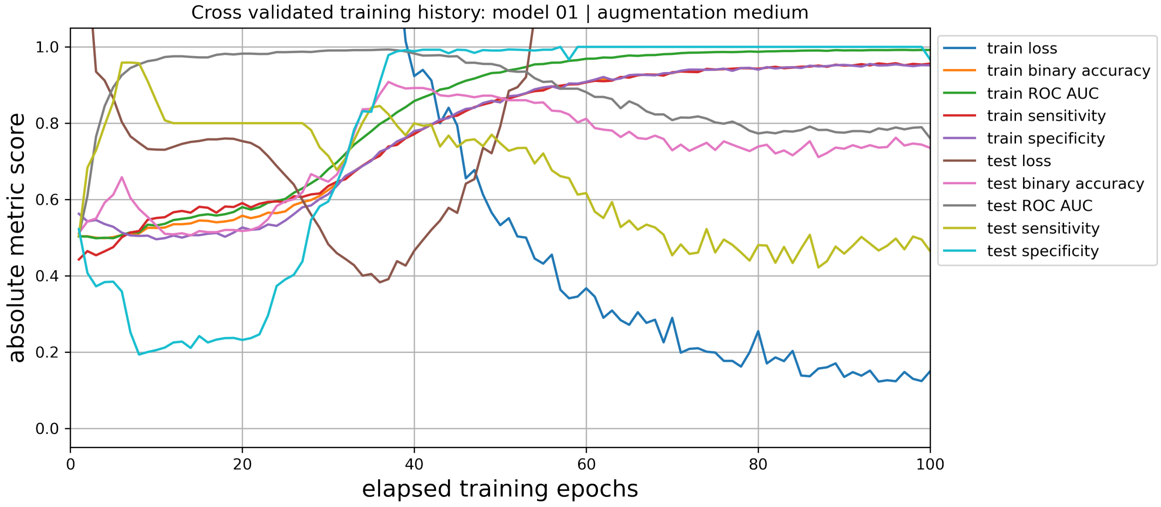


Figure 6.5. Quality scores history for dataset with augmentation medium.

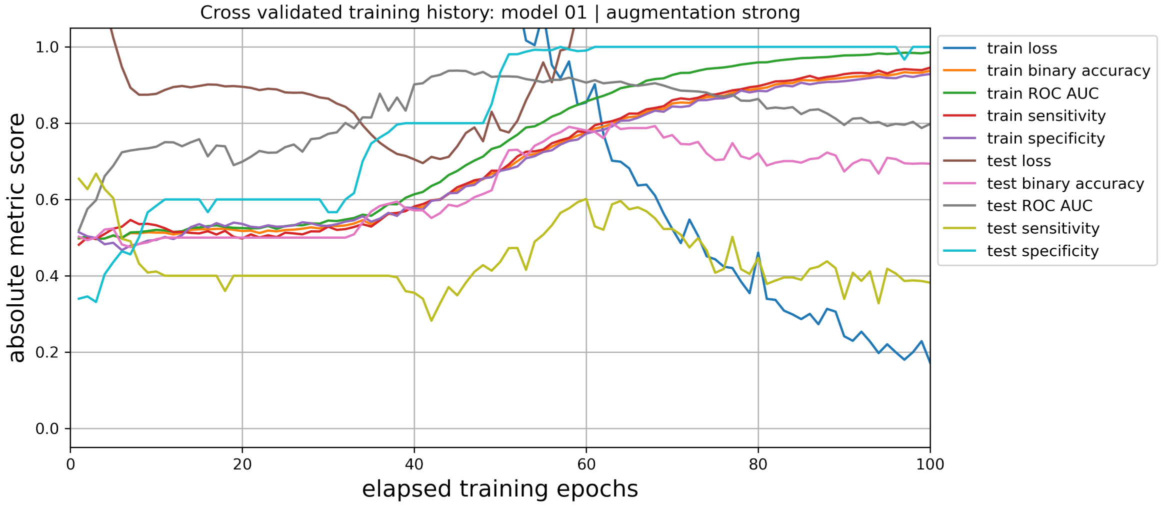


Figure 6.6. Quality scores history for dataset with augmentation strong.

The best epoch for the models consuming each dataset was determined based on the highest binary accuracy score over the entire training epochs, therefore, the remaining associated scores are the ones obtained exactly at that same epoch, rather than the maximum obtained across all epochs. Further, the dataset augmented with the easy level promoted the best performance overall, despite being outperformed in absolute scores by the medium augmentation level, since the cross validation process yielded far steadier results for the first one. E. g., despite the recall being 0.7602 and 0.8379 for augmentations easy and medium, respectively, the confidence interval with 99% probability for them were too discrepant being 0.1798 and 0.3041, respectively.

Table 6.1. 5-fold cross validation results for easy data augmentation level.

		Binary Accuracy	ROC AUC	Precision	Recall	F1	Specificity
Cross Validation Metric Property	Mean	0.8801	0.9916	1.0000	0.7602	0.8551	1.0000
	Standard Deviation	0.0779	0.0117	0.0000	0.1558	0.0985	0.0000
	Variance	0.0061	0.0001	0.0000	0.0243	0.0097	0.0000
	Confidence Interval 95%	0.0683	0.0102	0.0000	0.1366	0.0863	0.0000
	Confidence Interval 99%	0.0899	0.0135	0.0000	0.1798	0.1136	0.0000

Table 6.2. 5-fold cross validation results for light data augmentation level.

		Binary Accuracy	ROC AUC	Precision	Recall	F1	Specificity
Cross Validation Metric Property	Mean	0.8470	0.9752	0.8000	0.6940	0.7363	1.0000
	Standard Deviation	0.1877	0.0300	0.4000	0.3754	0.3792	0.0000
	Variance	0.0352	0.0009	0.1600	0.1409	0.1438	0.0000
	Confidence Interval 95%	0.1645	0.0263	0.3506	0.3290	0.3324	0.0000
	Confidence Interval 99%	0.2165	0.0346	0.4615	0.4331	0.4375	0.0000

Table 6.3. 5-fold cross validation results for medium data augmentation level.

		Binary Accuracy	ROC AUC	Precision	Recall	F1	Specificity
Cross Validation Metric Property	Mean	0.9082	0.9913	0.9793	0.8379	0.8724	0.9786
	Standard Deviation	0.1274	0.0156	0.0310	0.2636	0.1983	0.0323
	Variance	0.0162	0.0002	0.0010	0.0695	0.0393	0.0010
	Confidence Interval 95%	0.1117	0.0137	0.0272	0.2311	0.1738	0.0283
	Confidence Interval 99%	0.1470	0.0181	0.0358	0.3041	0.2288	0.0373

Table 6.4. 5-fold cross validation results for strong data augmentation level.

		Binary Accuracy	ROC AUC	Precision	Recall	F1	Specificity
Cross Validation Metric Property	Mean	0.7999	0.9731	1.0000	0.5998	0.7155	1.0000
	Standard Deviation	0.1321	0.0209	0.0000	0.2642	0.2093	0.0000
	Variance	0.0175	0.0004	0.0000	0.0698	0.0438	0.0000
	Confidence Interval 95%	0.1158	0.0183	0.0000	0.2316	0.1835	0.0000
	Confidence Interval 99%	0.1524	0.0241	0.0000	0.3049	0.2415	0.0000

6.3 Pain Samples Incorrectly Predicted Along the 10 Seconds Window

In pursuance of understanding the evolution of fetal responses to nociceptive input we have assessed using machine learning the difficulty of detecting pain along the 10 seconds time window immediately after the acute pin prick for the Pain group. Figure 6.7 presents the normalized cumulative incorrect predictions of painful images, averaged with cross validation, as a metric to express the difficulty of detecting pain as a function of time after the beginning of the painful stimulus. For each of the 5 folds in the cross validation, the painful frames have been organized in the 10 seconds time line, the relative error rate for each fold has been computed resulting in a value in the $[0, 1]$ interval, then the average among all folds has been taken. For the special cases where no image was available on a specific fold for a specific time interval, the cross validation was averaged with a smaller denominator. Finally, the resulting cross validated real numbers range in $[0, 1]$, however, each time point aggregates all the previous with itself, ultimately leading to bins with values potentially greater than 1. The chart is presented as a cumulative sum in order to become easier to understand and explain it.

Further, the red curve illustrates the changes in the height of bins, and it is formed by the aggregation of first degree functions between any consecutive bins. Between consecutive bins and above the red curve the slope of the function is presented, it represents the difficulty increase of correctly predicting images belonging to a further time window. The slope between bins may be compared with one another to determine which increases the most the degree of difficult to correctly detect pain.

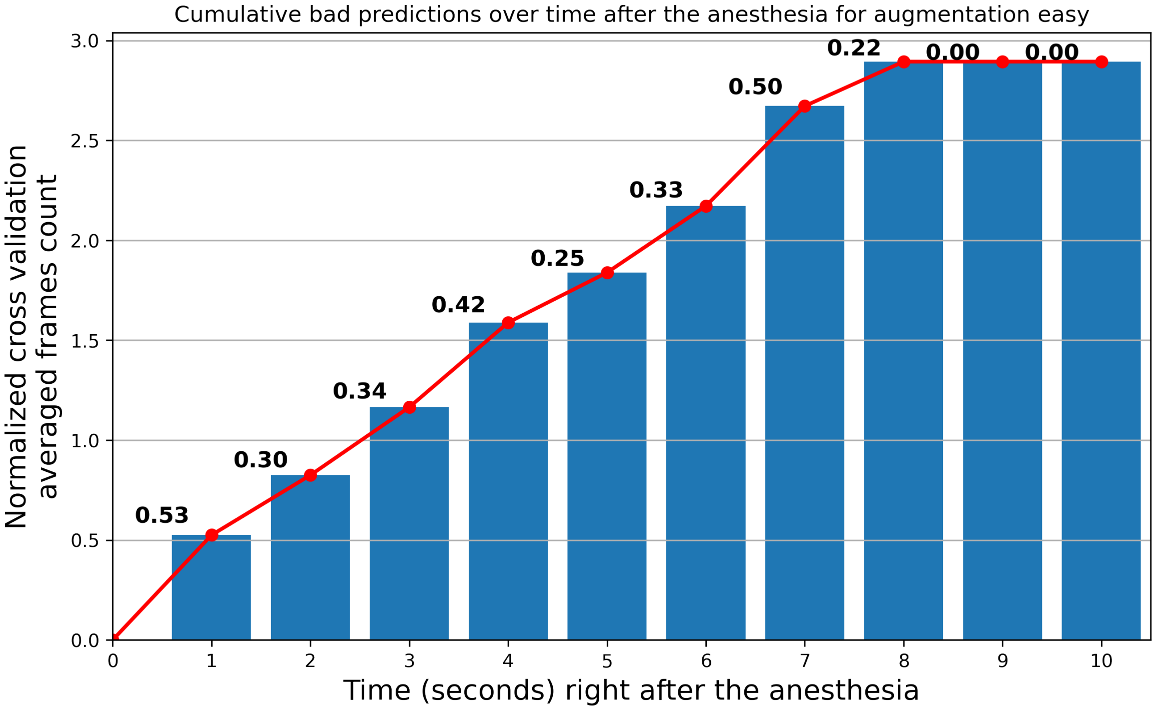


Figure 6.7. Evolution of the difficulty to detect pain along 10 seconds immediately after the beginning of the stimulus for the easy augmentation level dataset.

Table 6.5. Bad predictions of painful facial expressions along time for augmentation level easy. On each cell, the rightmost number represents the total amount of images on its own time interval as a fraction denominator, while the leftmost number represents the amount of images incorrectly predicted by our models as a fraction numerator. Both cross validation average metrics were computed only among folds that have at least one image in a given time interval, therefore, cells without any frames were filled with NaNs (Not a Number) and their values were not considered in the final average for that particular time interval.

		Time Window (Seconds)									
		[0, 1[	[1, 2[	[2, 3[	[3, 4[	[4, 5[	[5, 6[	[6, 7[	[7, 8[	[8, 9[	[9, 10]
Cross Validation Folds	0	1/4	0/1	1/5	3/4	2/2	4/4	2/2	0/3	0/5	0/5
	1	1/1	1/2	0/1	0/3	0/3	0/2	1/2	0/2	0/4	0/4
	2	3/8	0/10	5/10	4/9	0/5	NaN	NaN	NaN	NaN	NaN
	3	0/6	0/9	0/1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	4	1/1	2/2	3/3	2/4	0/1	0/4	0/3	2/3	0/3	NaN
Non Cumulative Cross Validation Average		0.5250	0.3000	0.3400	0.4236	0.2500	0.3333	0.5000	0.2222	0.0000	0.0000
Cumulative Cross Validation Average		0.5250	0.8250	1.1650	1.5886	1.8386	2.1719	2.6719	2.8942	2.8942	2.8942

6.4 Explainability Evaluation

This section approaches the XAI (eXplainable Artificial Intelligence) introduced in Section 4.12, along with the algorithms from Section 5.7, to generate explanations for the testing group of each fold of the dataset generated with the easy data augmentation level. This dataset was elected the best according to the results presented by Sectio 6.2, therefore, no XAI will be presented to the remaining ones despite the fact that they were generated by the experiments. Similarly to the idea proposed in Section 5.6, the testing groups were targeted since they weren't leaked into their respective training groups, in addition to not having been augmented. Had these factors been different, the XAI results could have become biased.

Furthermore, Figure 6.8 presents the *jet* colors scale used for the heatmaps on the XAI samples, that is Figures 6.9, 6.10, 6.11, 6.12, 6.13, 6.14, 6.15, 6.16, 6.17 and 6.18. Essentially, the color scale presents how intensely each region of pixel(s) in the target image contributed towards the ML models' produced answer. Therefore, note that the contribution level is not particularly related to the presence or absence of pain, it is actually related to the particular output the model produced for the respective image sample.

On this matter, the XAI samples were grouped according to the fold they belong to, in addition to the label they have. The original samples were presented in colors to serve as reference, however, its corresponding explained versions were converted to gray scale and the XAI heatmaps were placed on top as a semitransparent overlay. Each row represent a single image sample and each column represent a version of it, being the first the original image, while the remaining ones its corresponding XAI version generated by a particular algorithm, as described on top of each column. On the left hand side of those figures, each row title identifies whether the corresponding image label was correctly or incorrectly predicted.

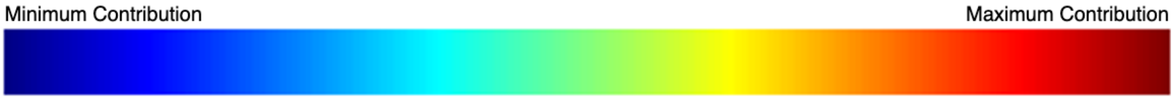


Figure 6.8. Explainability heatmap scale, it defines with colors how impactful regions over the target image were for the models' decision. This set of colors are used as an overlay representing a heatmap for each of the explainability images, i.e., Figures 6.9, 6.10, 6.11, 6.12, 6.13, 6.14, 6.15, 6.16, 6.17 and 6.18.

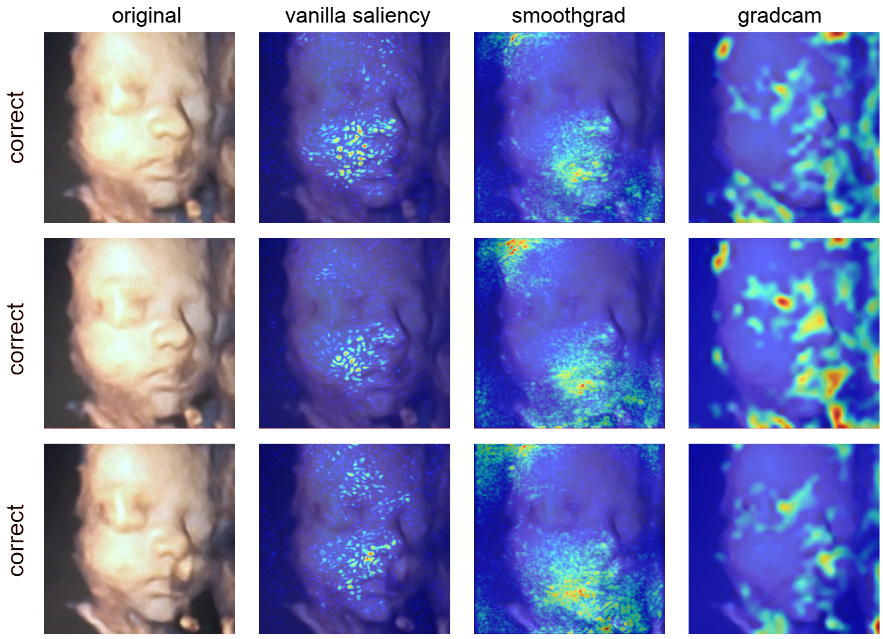


Figure 6.9. Original non pain image samples from fold 0 test set explained.

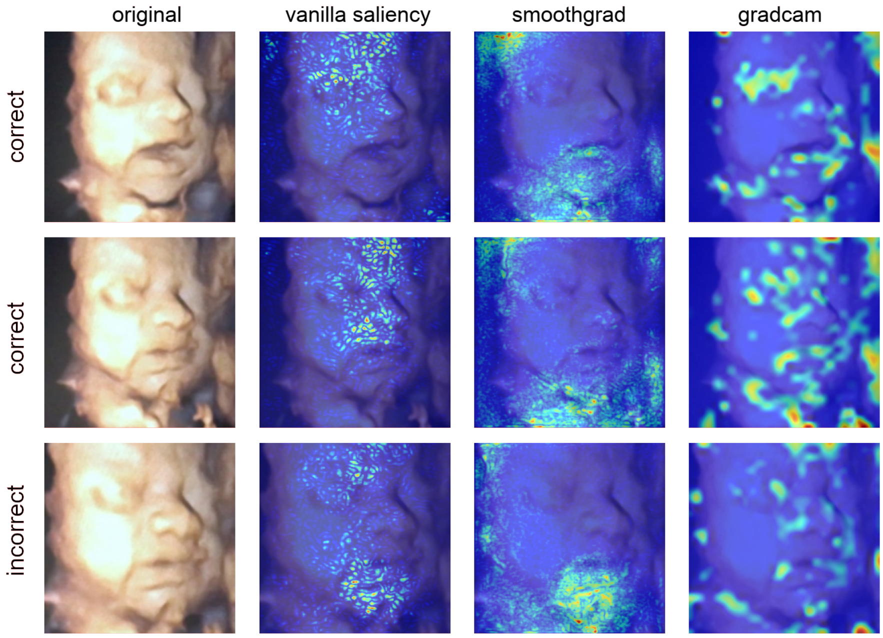


Figure 6.10. Original pain image samples from fold 0 test set explained.

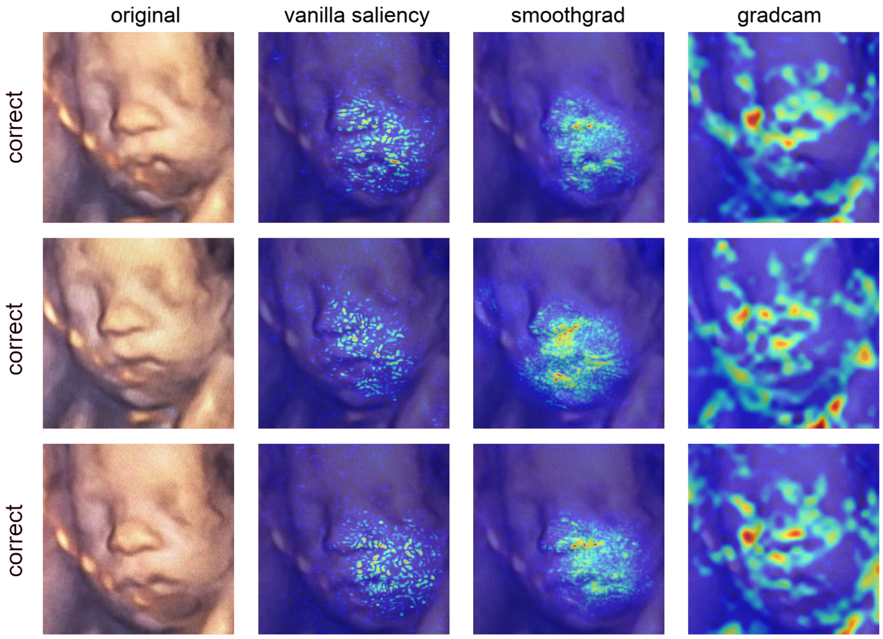


Figure 6.11. Original non pain image samples from fold 1 test set explained.

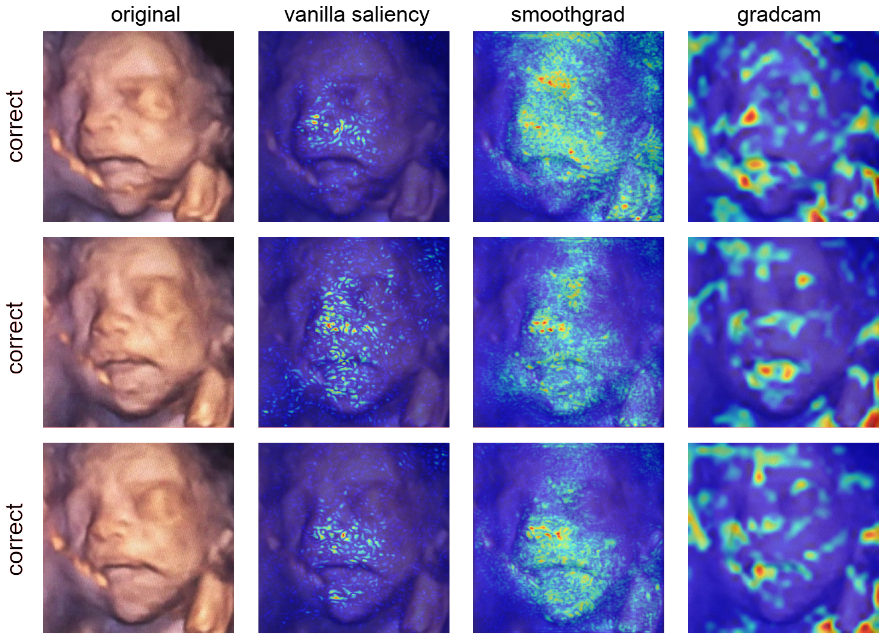


Figure 6.12. Original pain image samples from fold 1 test set explained.

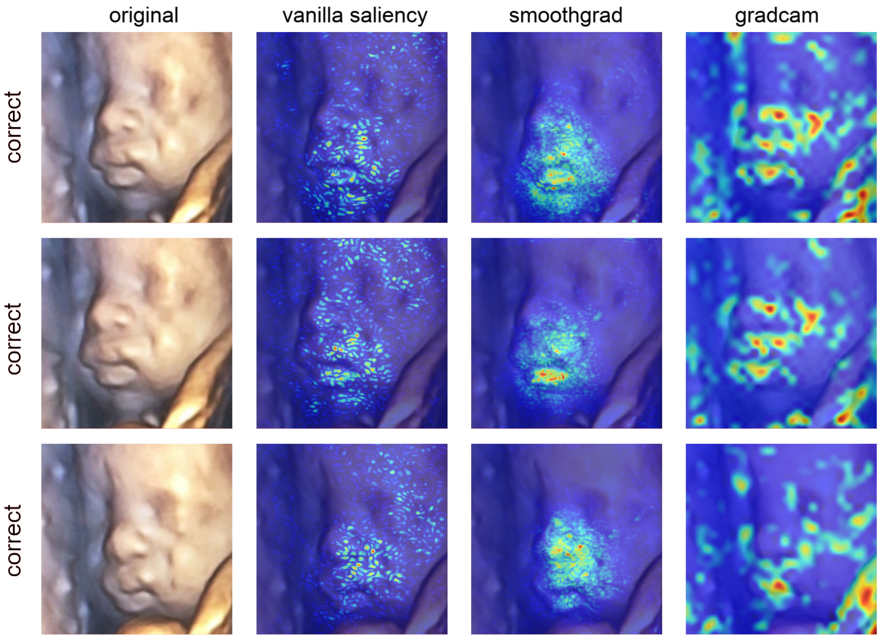


Figure 6.13. Original non pain image samples from fold 2 test set explained.

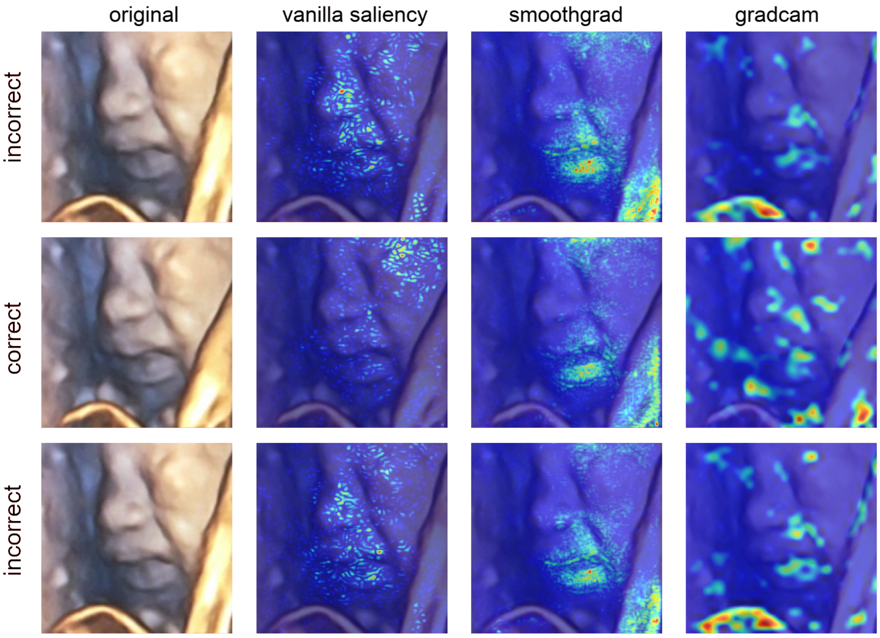


Figure 6.14. Original pain image samples from fold 2 test set explained.

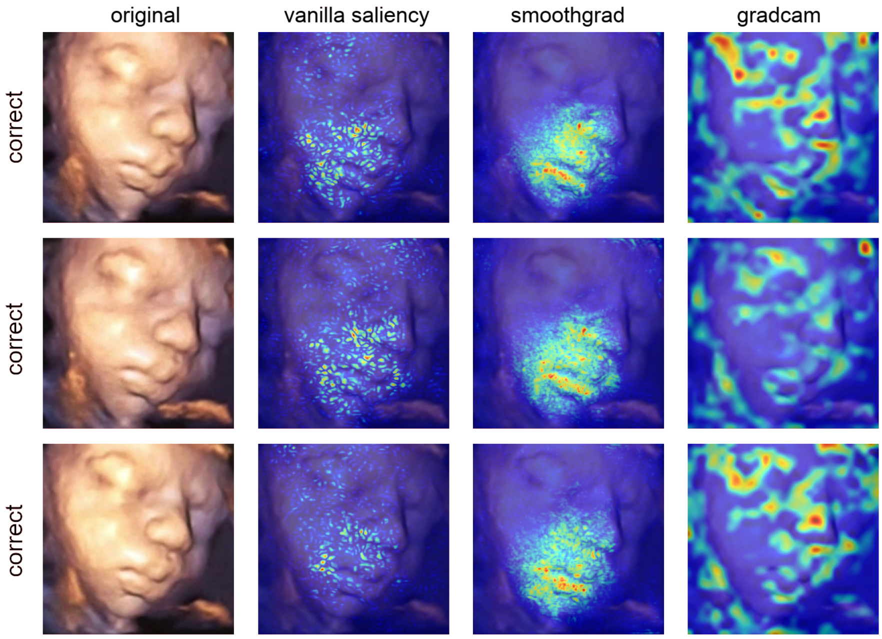


Figure 6.15. Original non pain image samples from fold 3 test set explained.

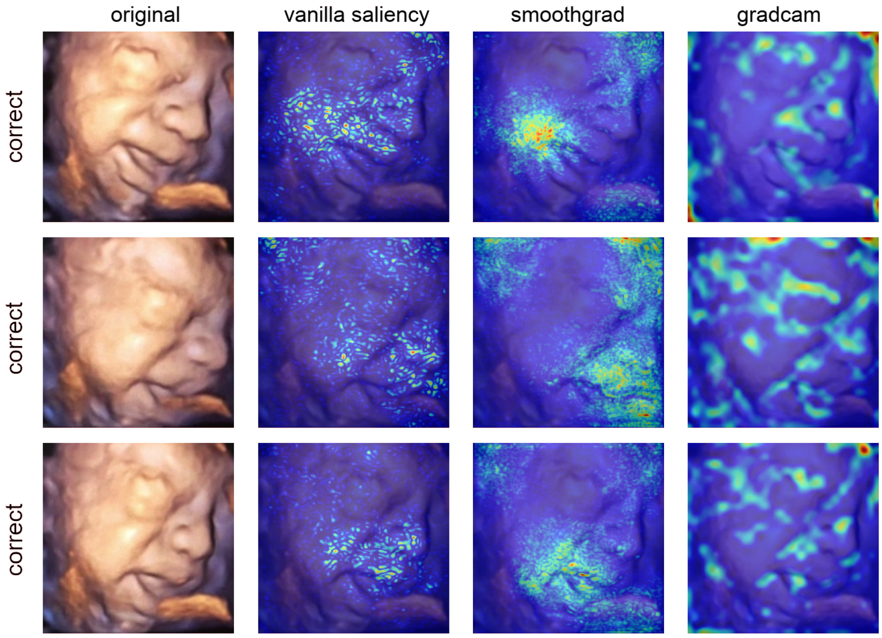


Figure 6.16. Original pain image samples from fold 3 test set explained.

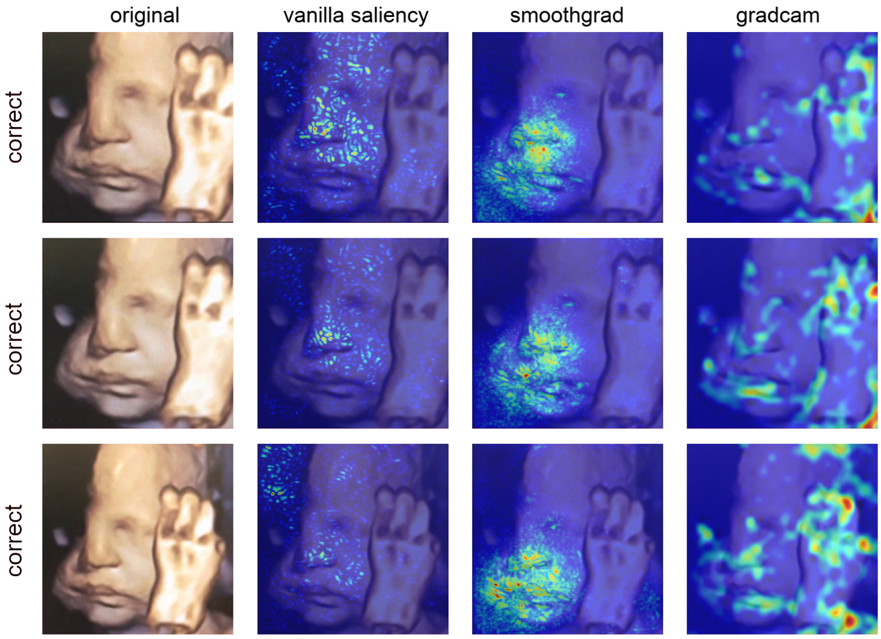


Figure 6.17. Original non pain image samples from fold 4 test set explained.

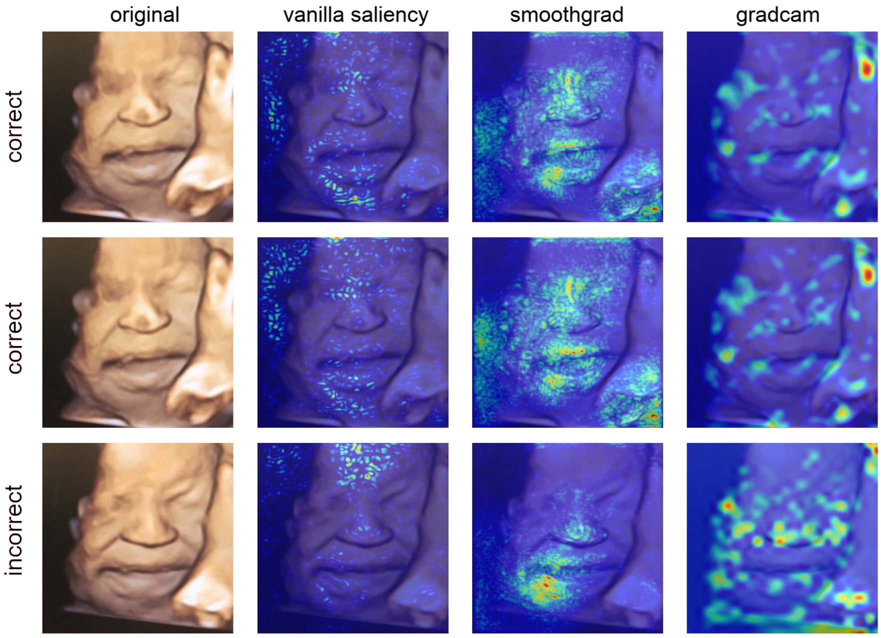


Figure 6.18. Original pain image samples from fold 4 test set explained.

Chapter 7

Conclusion

The cross validation technique was employed as a method to mitigate the uncertainty intrinsically present in the resulting metrics, however, their variance and standard deviation were fairly high. An important reason for this lies in the fact that the latent space provided as output by the VGGF16 feature extraction component has a huge dimensional space, with approximately 100 thousand and 25 thousand elements for the 4^{th} and 5^{th} maximum pooling layers. Such high dimensional space unbalances the tradeoff between bias and variance in favor of the last one [Hastie et al., 2009], something that could be avoided in case the VGGF16 convolutional component wasn't added to the models. The results were great despite the high standard deviation and broad confidence intervals, since obtaining better performance for complex tasks is extremely difficult on low diversity datasets such as the ones used in this project. In addition, the XAI results reinforce the importance of already known items present in the fetal-5 scale, leading to the conclusion that the models were indeed capable of generalizing well.

From the perspective that the trained AI models are tools for testing whether a fetus is in pain, it is really important to consider the obtained sensitivity and specificity. The first one measures the capacity of a test to correctly identify positive cases, while the second one to correctly identify negative cases. Mistakenly providing analgesic treatment for fetuses that aren't under pain isn't ideal, however, leaving unattended fetuses under pain is severe. The clear distinction between the two scenarios present the sensitivity metric as the most critical, therefore, a good fetal pain test would have high sensitivity.

The dropout was certainly a necessary regularizer, given that we had available few training samples for such complex problem to solve. Therefore, detecting the correct patterns which enable strong generalization power to the models ultimately leads to avoiding overfitting by slowing the pace at which the trainable parameters and the loss function changes, in addition to facilitating the model to learn multiple paths which lead to impactful results. Nonetheless, the attention technique was also extremely important to learn the correct association between segmented facial features which are, indeed, related to the presence or absence of pain. Painful facial expressions are usually comprised of multiple artifacts such as brow lowering or eyes squeezing shut [Bernardes et al., 2021], for which their co-occurrence increases the likelihood for the presence of pain, therefore, the attention layer in the neural network model is important to enable its internal learning of how intensely correlated such elements are, improving the quality of the final output probability. Further, attention commonly works well with embeddings, something similar to the output provided by the VGGF16 feature extractor component. Moreover, the ELU activation function was also critical, mostly due to its capabilities of mitigating the vanishing gradient problem via the identity for positive values.

Further, as shown by Figure 6.7, it is difficult to correctly detect pain in the first seconds after the begin of the painful stimulation. We have considered that right on the painful stimulus and 10 seconds onwards every image will present pain, however, as shown by the evolution of the difficulty to detect pain along time, the very beginning seems to be the most challenging, while the final seconds seems to be the easiest. Perhaps the reaction time of the fetuses to express pain is impactful, however, we considered everything right after the beginning of the stimulus as pain. In addition, fetuses from the Acute Pain group pre-stimulus were inserted in the Non-Pain group, however, they were suffering from an illness. Moreover, similar to the Acute Pain scenario, samples from the Co-AS group immediately after the acoustic stimulation were scarce. All these factors have the potential to create biases that could affect the obtained results, nonetheless, they were mitigated whenever possible.

The tests executed in this project consisted of a single chance for the ML models to predict the correct label of each image instance, however, in the real world, a single 4D-US assessment of a fetus in video would generate plenty of images. The models would have a much better estimate of pain on a particular fetus in case of post processing the inference results for all those image samples. The same fetus could be assessed through different angles and positions with multiple chances to predict the correct answer, something that would increase the quality of the estimations, ultimately leading to increased sensitivity and specificity.

On the matter of the XAI results presented in Section 6.4, the vanilla saliency and smoothgrad explanation algorithms present fine-grained regions of the target image as relevant, and there were cases where these regions also clumped up together to form broader ones. By another hand, gradcam presented coarser-grained regions as impactful when compared to the first two algorithms, however, it was traced back to the VGGF16 layer 14, instead of the 18th, and as a result the impacful regions became finer-grained than usual. Nonetheless, there is a clear trend presented by all 3 XAI algorithms, indicating that facial elements are important to produce the correct output. Further, such trend presents very frequently the regions comprehending the fetuses chin, mouth, cheeks and nose as the most relevant, something directly related to known items from the fetal-5 scale. Despite the resulting metrics, the explanations make clear that the ML models did manage to generalize well, by capturing such facial traits, ultimately increasing the confidence we may have on the results. According to Bernardes et al. [2021], mouth related items such as horizontal and vertical stretching or open lips are the facial items that best discriminates the presence of pain, something highlighted by the XAI results. Due to the mathematical optimization that consists the training process, ML models tend to learn characteristics with as much discriminative capabilities as they can over the training data, therefore, XAI results have confirmed how intensely discriminative the fetal mouth can be. Further, more items from the fetal-5 scale were also pointed as relevant, such as the regions from the eyes or the nasolabial furrow, however, artifacts absent in the fetal-5 scale were also highlighted, such as arms, or wrinkles across the fetal face in addition to the uterus internal wall, among more. Confirming whether items outside the fetal-5 scale does have relevant contributions towards pain detection requires further investigation, however, the mouth and surrounding regions on the fetal face remains as the most impactful ones.

In order to become a feasible product in the market, it is necessary to eliminate any manual steps of preprocessing employed. Images used to train and test the ML models were manually cropped to focus on the fetal face, something that could be automated by hand-crafting another dataset where another ML model would determine the bounding box surrounding the target area as a regression task. Similarly to what is done by Carion et al. [2020], this project could benefit from the detection of multiple fetal faces, in case of multiple fetuses in a single pregnancy, however, the transformer based model architecture might have greater capacity than the amount of training data available, implicating a shallower model might lead to better results.

Chapter 8

Future Work

In order to enable the present work to cause positive impact in this world, it would be necessary to turn it into a product for the market, as already suggested in Chapter 7. The main purpose of creating an automated pain testing tool is to empower physicians and caregivers on the assessment of the fetal context, rather than substituting them. Further, as 4D-US machines become more portable and accessible, non medical professionals may evaluate the pregnancy at home and seek professional assistance in case the proposed ML model outputs high likelihood for pain.

Moreover, as suggested by Chapter 7, investigating whether some artifacts detected by our ML models and highlighted by the corresponding XAI algorithms remains an open issue. Despite the conclusion that the fetal mouth is critical to discriminate pain, it isn't yet possible to confirm whether the same is true for other detected artifacts.

Further, we currently plan to apply the Fetal-5 score to a large number of pregnancies (n=200) and assess to which extent it can add to the current fetal vitality parameters routinely used in clinical practice, i.e., pain assessment is important to address general fetal vitality. In addition, the methodology developed in this project for assessing pain is similar to the one expected to be employed on the fetal vitality assessment, making it a natural continuation of this project.

Bibliography

AboEllail, M. A. M. and Hata, T. (2017). Fetal face as important indicator of fetal brain function. *Journal of Perinatal Medicine*, 45(6):729--736.

Adzick, N. S., Thom, E. A., Spong, C. Y., Brock, 3rd, J. W., Burrows, P. K., Johnson, M. P., Howell, L. J., Farrell, J. A., Dabrowiak, M. E., Sutton, L. N., Gupta, N., Tulipan, N. B., D'Alton, M. E., Farmer, D. L., and MOMS Investigators (2011). A randomized trial of prenatal versus postnatal repair of myelomeningocele. *New England Journal of Medicine*, 364(11):993--1004.

Apgar, V. (1966). The newborn (apgar) scoring system: Reflections and advice. *Pediatric Clinics of North America*, 13(3):645–650. The Newborn I.

Barr, R. G., Hopkins, B., and Green, J. A., editors (2000). *Crying as a Sign, a Symptom, and a Signal.* Cambridge University press, Mac Keith Press.

Bearak, J., Popinchalk, A., Ganatra, B., Moller, A.-B., Tuncalp, O., Beavin, C., Kwok, L., and Alkema, L. (2020). Unintended pregnancy and abortion by income, region, and the legal status of abortion: estimates from a comprehensive model for 1990-2019. *The Lancet Global Health*, 8:E1152--E1161.

Bellieni, C. V. (2012). Pain assessment in human fetus and infants. *The American Association of Pharmaceutical Scientists journal*, 14(3):456--461.

Bernardes, L. S., Carvalho, M. A., Harnik, S. B., Teixeira, M. J., Ottolia, J., Castro, D., Veloso, A., Francisco, R., Listik, C., Galhardoni, R., da Silva, V. A., Moreira, L. I., de Amorim Filho, A. G., Fernandes, A. M., and de Andrade, D. C. (2021). Sorting pain out of salience: assessment of pain facial expressions in the human fetus. *Pain Reports*, 6(1)(e882).

Bernardes, L. S., Ottolia, J. F., Cecchini, M., de Amorim Filho, A. G., Teixeira, M. J., Francisco, R. P. V., de Andrade, D. C., and de Estudo da Dor Fetal (Fetal Pain

Study Group), G. (2018). On the feasibility of accessing acute pain-related facial expressions in the human fetus and its potential implications: a case report. *Pain Reports*, 3(5)(e673).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J., editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213--229. Springer.

Church, C. C. and Miller, M. W. (2007). Quantification of risk from fetal exposure to diagnostic ultrasound. *Progress in Biophysics and Molecular Biology*, 93(1):331–353. Effects of ultrasound and infrasound relevant to human health.

Clevert, D., Unterthiner, T., and Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Craig, K. D., Whitfield, M. F., Grunau, R. V. E., Linton, J., and Hadjistavropoulos, H. D. (1993). Pain in the preterm neonate: behavioural and physiological indices. *Pain*, 52(3):287--299.

de Graaf-Peters, V. B. and Hadders-Algra, M. (2006). Ontogeny of the human central nervous system: what is happening when? *Early Human Development*, 82(4):257--266.

de Oliveira, T. M., Veloso, A., and Ziviani, N. (2019). Automatic pain assessment in fetuses through transfer learning. Master's thesis, Universidade Federal de Minas Gerais.

Debillon, T., Zupan, V., Ravault, N., Magny, J. F., and Dehan, M. (2001). Development and initial validation of the EDIN scale, a new tool for assessing prolonged pain in preterm infants. *Archives of Disease in Childhood: Fetal & Neonatal*, 85(1):F36--41.

Fantz, A. (2016). Utah passes new abortion law. `https://edition.cnn.com/2016/03/29/health/utah-abortion-law-fetal-pain/index.html`. Accessed: 2021-12-10.

Finley, G. A. and McGrath, P. J. (1998). Measurement of pain in infants and children. *Journal of Pediatric Hematology/Oncology*, 20:364.

Fitzgerald, M. (1987). Prenatal growth of fine-diameter primary afferents into the rat spinal cord: a transganglionic tracer study. *Journal of Comparative Neurology*, 261(1):98--104.

Giannakoulopoulos, X., Sepulveda, W., Kourtis, P., Glover, V., and Fisk, N. M. (1994). Fetal plasma cortisol and beta-endorphin response to intrauterine needling. *Lancet*, 344(8915):77--81.

Giannakoulopoulos, X., Teixeira, J., Fisk, N., and Glover, V. (1999). Human fetal and maternal noradrenaline responses to invasive procedures. *Pediatric Research*, 45(4 Pt 1):494--499.

Glover, V. and Fisk, N. M. (1999). Fetal pain: implications for research and practice. *British Journal of Obstetrics and Gynaecology*, 106(9):881--886.

Grichnik, K. P. and Ferrante, F. M. (1991). The difference between acute and chronic pain. *Mount Sinai Journal of Medicine*, 58(3):217--220.

Grunau, R. E., Oberlander, T., Holsti, L., and Whitfield, M. F. (1998). Bedside application of the neonatal facial coding system in pain assessment of premature neonates. *Pain*, 76(3):277--286.

Harris, C. G. and Stephens, M. (1988). A combined corner and edge detector. In Taylor, C. J., editor, *Proceedings of the Alvey Vision Conference, AVC 1988, Manchester, UK, September, 1988*, pages 1--6. Alvey Vision Club.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer.

Herrington, C. J., Olomu, I. N., and Geller, S. M. (2004). Salivary cortisol as indicators of pain in preterm infants: a pilot study. *Clinical Nursing Research*, 13(1):53--68.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *Computing Research Repository*, abs/1207.0580 `http://arxiv.org/abs/1207.0580`.

Houtrow, A. J., Thom, E. A., Fletcher, J. M., Burrows, P. K., Adzick, N. S., Thomas, N. H., Brock, John W., I., Cooper, T., Lee, H., Bilaniuk, L., Glenn, O. A., Pruthi, S., MacPherson, C., Farmer, D. L., Johnson, M. P., Howell, L. J., Gupta, N., and

Walker, W. O. (2020). Prenatal Repair of Myelomeningocele and School-age Functional Outcomes. *Pediatrics*, 145(2). e20191544.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*. `https://hal.inria.fr/inria-00321923/file/Huang_long_eccv2008-lfw.pdf`.

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery Data*, 6(4):15:1--15:21.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kolen, J. and Kremer, S. (2001). *A Field Guide to Dynamical Recurrent Networks*. Wiley.

Lawrence, J., Alcock, D., McGrath, P., Kay, J., MacMurray, S. B., and Dulberg, C. (1993). The development of a tool to assess neonatal pain. *Neonatal Network*, 12(6):59--66.

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In Forsyth, D. A., Mundy, J. L., Gesù, V. D., and Cipolla, R., editors, *Shape, Contour and Grouping in Computer Vision*, volume 1681 of *Lecture Notes in Computer Science*, page 319. Springer.

Mauricio, A., Cappabianco, F. A. M., Veloso, A., and Cámara, G. (2019). A sequential approach for pain recognition based on facial representations. In Tzovaras, D., Giakoumis, D., Vincze, M., and Argyros, A. A., editors, *Computer Vision Systems, 12th International Conference, ICVS 2019, Thessaloniki, Greece, September 23-25, 2019, Proceedings*, volume 11754 of *Lecture Notes in Computer Science*, pages 295--304. Springer.

Merskey, H. and Bogduk, N. (1994). Classification of chronic pain. descriptions of chronic pain syndromes and definitions of pain terms. *The Journal of the International Association for the Study of Pain*, 3:S1--226. Pain. Supplement 3.

Okado, N. (1981). Onset of synapse formation in the human spinal cord. *Journal of Comparative Neurology*, 201(2):211--219.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In Xie, X., Jones, M. W., and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference 2015, BMVC*, pages 41.1--41.12. BMVA Press.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1310--1318. JMLR.org.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *Computing Research Repository*, abs/1712.04621 `http://arxiv.org/abs/1712.04621`.

Peters, J. W. B., Koot, H. M., Grunau, R. E., de Boer, J., van Druenen, M. J., Tibboel, D., and Duivenvoorden, H. J. (2003). Neonatal facial coding system for assessing postoperative pain in infants: item reduction is valid and feasible. *Clinical Journal of Pain*, 19(6):353--363.

Rakitianskaia, A. S. and Engelbrecht, A. P. (2015). Measuring saturation in neural networks. In *IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015*, pages 1423--1430. IEEE.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336--359.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Slater, R., Cantarella, A., Gallella, S., Worley, A., Boyd, S., Meek, J., and Fitzgerald, M. (2006). Cortical pain responses in human infants. *J. Neurosci.*, 26(14):3662--3666.

Slater, R., Worley, A., Fabrizi, L., Roberts, S., Meek, J., Boyd, S., and Fitzgerald, M. (2010). Evoked potentials generated by noxious stimulation in the human infant brain. *European Journal of Pain*, 14(3):321--326.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *Computing Research Repository*, abs/1706.03825 `http://arxiv.org/abs/1706.03825`.

Smith, S. L., Dherin, B., Barrett, D. G. T., and De, S. (2021). On the origin of implicit regularization in stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929--1958.

Stevens, B., Johnston, C., Petryshen, P., and Taddio, A. (1996). Premature infant pain profile: development and initial validation. *Clinical Journal of Pain*, 12(1):13--22.

Suraseranivongse, S., Kaosaard, R., Intakong, P., Pornsiriprasert, S., Karnchana, Y., Kaopinpruck, J., and Sangjeen, K. (2006). A comparison of postoperative pain scales in neonates. *British Journal of Anaesthesia*, 97(4):540--544.

Van de Velde, M. and De Buck, F. (2012). Fetal and maternal analgesia/anesthesia for fetal procedures. *Fetal Diagnosis and Therapy*, 31(4):201--209.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998--6008.

Wolf, L., Hassner, T., and Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 529--534. IEEE Computer Society.

Zamzmi, G., Goldgof, D. B., Kasturi, R., and Sun, Y. (2018). Neonatal pain expression recognition using transfer learning. *Computing Research Repository*, abs/1807.01631 `http://arxiv.org/abs/1807.01631`.

Zamzmi, G., Pai, C., Goldgof, D. B., Kasturi, R., Sun, Y., and Ashmeade, T. (2016). Machine-based multimodal pain assessment tool for infants: A review. *Computing Research Repository*, abs/1607.00331 `http://arxiv.org/abs/1607.00331`.

Zamzmi, G., Pai, C., Goldgof, D. B., Kasturi, R., Sun, Y., and Ashmeade, T. (2017). Automated pain assessment in neonates. In Sharma, P. and Bianchi, F. M., editors, *Image Analysis - 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12-14, 2017, Proceedings, Part II*, volume 10270 of *Lecture Notes in Computer Science*, pages 350--361. Springer.

Zamzmi, G., Paul, R., Salekin, M. S., Goldgof, D. B., Kasturi, R., Ho, T., and Sun, Y. (2019). Convolutional neural networks for neonatal pain assessment. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(3):192--200.

Zimmermann, M. (1991). Pain in the fetus: neurobiological, psychophysiological and behavioral aspects. *Schmerz*, 5(3):122--130.

Attachment A

Glossary

- AI: Artificial Intelligence;
- AP (group of fetuses): Acute Pain [Bernardes et al., 2018, 2021];
- Acute pain (health condition): is provoked by a specific disease or injury, serves a useful biologic purpose, is associated with skeletal muscle spasm and sympathetic nervous system activation, and is self-limited [Grichnik and Ferrante, 1991];
- ANNs: Artificial Neural Networks;
- Co-Re (group of fetuses): Control at Rest [Bernardes et al., 2018, 2021];
- Co-AS (group of fetuses): Control Acoustic Startle [Bernardes et al., 2018, 2021];
- Co-Re-AS (group of fetuses): Control at Rest and Acoustic Startle [Bernardes et al., 2018, 2021];
- Chronic pain (health condition): may be considered a disease state. It is pain that outlasts the normal time of healing, if associated with a disease or injury. Chronic pain may arise from psychological states, serves no biologic purpose, and has no recognizable end-point [Grichnik and Ferrante, 1991];
- CNNs: Convolutional Neural Networks;
- DA: Data Augmentation;
- DL: Deep Learning;
- FPSG: Fetal Pain Study Group;
- IASP: International Association for the Study of Pain;

- ML: Machine Learning;
- MOMS: Management Of Myelomeningocele Study;
- NaN: Not a Number, represents an invalid value, generally created with operations such as division by zero;
- NFCS: Neonatal Facial Coding System;
- NNs: Neural Networks (another name for ANNs when the term artificial is implied);
- PDEF: Partial Derivative of the Error Function;
- ResNets: Residual (Neural) Networks;
- ROC AUC: Receiver Operating Characteristic Area Under the Curve;
- SOTA: State-Of-The-Art, refers to the most advanced stage in the development of something such as a new technology or product;
- TL: Transfer Learning;
- USP: *Universidade de São Paulo*, a university located in the city of São Paulo-SP Brazil;
- VGG (research group): Visual Geometry Group, is an academic group focused on computer vision at Oxford University;
- VGGF16: VGG-Face-16, a very deep convolutional neural network architecture;
- XAI: eXplainable Artificial Intelligence;