A Convolutional Neural Network-based Mobile Application to Bedside Neonatal Pain Assessment

Lucas P. Carlini[1], Leonardo A. Ferreira[1], Gabriel A. S. Coutrin[1],
Victor V. Varoto[1], Tatiany M. Heiderich[1],
Rita C. X. Balda[2], Marina C. M. Barros[2], Ruth Guinsburg[2], Carlos E. Thomaz[1]
[1]University Center of FEI, São Bernardo do Campo SP, Brazil
[2]Federal University of São Paulo, São Paulo SP, Brazil
lucaspcarlini10,ruth.guinsburg@gmail.com, cet@fei.edu.br

***Abstract*—More than 500 painful interventions are carried out during the hospitalisation of a newborn baby in a neonatal intensive care unit. Since neonates are not able to verbally communicate pain, some studies have been done to identify the presence and intensity of pain by behavioural analysis, mainly by facial expression. These studies allow a better understanding of this painful experience faced by the neonate. In this context, this work proposes and implements a mobile application for smartphones that uses Artificial Intelligence (AI) techniques to automatically identify the facial expression of pain in neonates, presenting feasibility in real clinical situations. Firstly, a Convolutional Neural Network architecture was adapted and trained with face images captured before and after painful clinical procedures carried out routinely. Then, this computational model was optimised to a mobile environment to make it practical for everyday use. Moreover, we used an explainable AI method to identify facial regions that might be relevant to pain assessment. Our results showed that is possible to classify the facial expression of the pain of neonates with high accuracy. Additionally, our methodology presented novel results highlighting as well sound facial regions that agree with pain scales used by neonatologists and with the visual perception of adults when assessing pain in neonates, whether they are health professionals or not.**

I. Introduction

More than 500 painful interventions are carried out during the hospitalisation of a newborn baby in a neonatal intensive care unit [1], [2]. In the past, due to the inability of neonates to verbally communicate pain, it was believed that the central nervous system of newborn babies was not fully developed, consequently, not being able to sense and suffer from pain [3]. However, it was observed in the latter years of 1980 that the central nervous and nociceptive systems are sufficiently developed in the sixth month of gestation [4], [5], leading to an increased sensitivity of pain, since inhibitory pathways of painful stimulus are not fully developed yet [6], [7].

The non-treated pain felt by neonates is associated with changes in their respiratory, cardiovascular and metabolic stability, increasing mortality in neonatal intensive care units. In the short-term, neonates may suffer, as a consequence of pain, irritability, inattention, change in resting pattern, dietary denial, and interference in the mother-child relationship [3]. In the long-term, pain may cause degrading effects on the neurological and behavioural development such as cognitive problems, alterations in brain development with implications on learning disabilities, and hypersensitivity to painful and non-painful stimuli [1], [8].

The most common observations reported by health professionals when treating neonatal pain are crying, irritability, sudden movements, and change in their facial expression and behaviour [9]. Due to the range of sounds, most people believe that crying is the best expression to estimate the presence and, if necessary, the intensity of pain or discomfort in their newborn babies. However, it has been argued that almost half of newborn babies do not cry in painful interventions, meanwhile, a stressful stimulus may lead to crying as well. Therefore, although useful, crying alone is not reliable to verify the presence of pain [10]. Another useful method to verify the presence of neonatal pain is the analysis of their movements. This can be justified by the fact that newborn babies have a standardised movement. Consequently, changes in that pattern, such as sudden and disorderly movements, may indicate the presence of pain. More recently, pain presence may be verified through the analysis of the facial expression of the neonate [3]. Although this analysis may be subjective, it is a noninvasive method that has been widely used in clinical practice that delivers valid information regarding the nature and intensity of the pain allowing better communication between the neonate and his/her caregiver [5], [10].

A. Related Work

Methods based on Artificial Intelligence (AI) and facial expression recognition have been proposed in the last 5 years that allow the implementation of non-invasive computational frameworks that are specific to the pain phenomena and enable continuous monitoring of the neonate.

In 2015, Heiderich et al. [3] proposed a computational framework that enables pain assessment through facial movements based on the NFCS [5], [10]. This framework was created using the Embarcadero Delphi XE2 software for Windows operating system. In order to measure the facial pattern of a neonate, Heiderich et al. used the LuxandFaceSDK software, enabling the identification of 66 facial landmarks, such as the contour of the eye, nose, and mouth, and the tip of the nose. When the neonate featured three or more of the facial actions based on the Neonatal Facial Coding System (NFCS) [5], [10],

it was considered that the neonate was suffering from pain. The results obtained by Heiderich et al. showed 85% of sensitivity and 100% specificity when assessing a neonate during periods of rest and 100% of sensitivity and specificity when assessing a neonate with pain. Among all the frameworks analysed in this section, this is the only one that was implemented to use in real clinical situations.

Later, Teruel et al. [11] implemented a pattern recognition and feature extraction method based on the framework proposed by Thomaz et al. [12]. This method uses Principal Component Analysis [13] in order to reduce the dimensionality of the input data (in our case, a neonate face image) and, then, uses the Maximum uncertainty Linear Discriminant Analysis [12] to identify the hyperplane that better discriminates the input data, enabling its classification as "pain" or "no pain". Teruel obtained an accuracy of 100% when comparing his results with the health-professionals classification. However, when comparing with the NFCS, the method showed 72.77% of accuracy.

In 2018, Zamzmi et al. proposed two different techniques based on the use of Convolutional Neural Networks (CNNs). Their first work [14] proposed the use of transfer learning to four pre-trained CNNs architectures: VGG-F, VGG-M, VGG-S, and VGG-Face [15]. VGG-F, M, S architectures were originally trained on the ImageNet dataset for object classification while VGG-Face was trained on a face-specific dataset. Deep features of each input data were extracted from high- and lower-layer of these architectures and, then, these features were used to train a supervised machine learning classifier. The results obtained achieved 0.841 AUC and 90.34% accuracy. It was also observed that the VGG-Face architecture performed better than others since this architecture was pre-trained on a face dataset, leading to a better feature extraction. More recently, Zamzmi et al. also proposed a Neonatal Convolutional Neural Network (N-CNN) [16], designed and trained end-to-end to detect neonatal pain. The architecture is a cascaded CNN that has three convolutional branches, allowing combining the image-specific information with the general information after applying convolutions. The features obtained by these branches are merged and, then, classified as "pain" or "no pain" by two fully connected layers. The proposed N-CNN achieved 91% average accuracy and 0.93 AUC on the Neonatal Pain Assessment Dataset [16] and 84.5% average accuracy on the infant Classification of Pain Expression dataset [17]–[19].

B. Contribution

This paper presents a novel computational framework based on CNNs to automatically identify the facial expression of pain in neonates, showing feasibility in real clinical situations. We approach this problem by applying transfer learning to a CNN architecture pre-trained with face images. We adapt this architecture adding fully connected layers specifically trained with neonatal face images captured before and after painful clinical procedures carried out routinely. Then, this classification model was embedded in a mobile application in order to make this framework practical for everyday use. Moreover, we used an explainable AI method to better understand the relationship between the classification model prediction in terms of its features (image pixels). To the best of our knowledge, this paper is the first to apply these techniques in neonatal pain classification, identifying facial regions that might be relevant to pain assessment, leading to a better comprehension of the facial expressiveness of a neonate when experiencing pain.

C. Outline

The paper is organised as follows. In Section II we describe the neonatal pain assessment datasets, computational methods, and implementation of our mobile application. Then, Section III presents evaluation metrics results and our findings on model interpretability. And Section IV comments on the feasibility of our mobile application. Finally, we conclude and provide guidance for further work in Section V.

II. Materials and Methods

This section is divided into three parts: (1) Face image datasets, (2) Computational methods, and (3) Implementation of our mobile application.

A. Face Image Datasets

We used two image datasets to design our proposed framework: UNIFESP [3] and iCOPE [17]–[19].

1) UNIFESP Image Dataset: The UNIFESP dataset was developed by Heiderich et al. [3] at the Federal University of São Paulo. It includes 30 healthy neonates (7 late preterms and 23 born at term) with 34 to 41 weeks of gestational age and 24 to 168 hours of life. For each neonate, it was recorded 10 minutes videos before, during, and after painful procedures. These procedures, such as venipuncture, capillary, or intramuscular injection, were performed while collecting routine tests or administering vaccines. After capturing the videos of each neonate, images were extracted every three seconds.

All photographs were captured using three Foscam cameras with a resolution of 320x233. From the images captured by the system, 12 images were chosen for each one of the 30 participating neonates. In total, 360 images were collected, of which: 138 were captured before a painful procedure, 30 during the procedure, and 192 images captured within 10 minutes after a painful procedure. Subsequently, each image was randomly submitted for evaluation by health professionals. These health professionals were neonatologists with experience working in neonatal intensive care units. The assessment led to 164 images classified as "in pain" and 196 images classified as "without pain".

It is noteworthy that these images have not undergone any kind of transformation or processing, maintaining the original features from the recording. Besides the face of the neonate, objects related to the hospitalisation and also other parts of the body, such as the neck and hands, are present in the images. Therefore, the similarity with real situations found in the intensive care units is preserved.

2) infant Classification of Pain Expression: The infant Classification of Pain Expression (iCOPE) dataset was developed by Brahnam et al. [17]–[19] during a study at the St. John Hospital (now called Mercy Hospital) with the Neonatology Department in Missouri, USA. A total of 200 images were captured from 26 neonates, 13 girls and 13 boys, all Caucasians. The age group of these neonates ranges from 18 hours to 3 days of life. Although all of them were in good health, it was reported that six male babies were circumcised the day before the photos were captured and that everyone's last feed was done in a period of 45 minutes to 5 hours before the photographs were taken.

All images were photographed using a Nikon D100 digital camera in ambient light conditions with a resolution of 3008x2000 in a room separate from other neonates. The neonates were photographed during a session in which they experienced 4 different stimuli performed in the following sequence:

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

This dataset is composed of: 63 images of neonates resting, 18 crying, 23 images of air stimulation, 36 during friction, and 60 with neonates during a painful procedure.

B. Computational Methods

To design our computational model, we performed three main steps:

1) Facial Detection: using a facial recognition algorithm, we extracted neonate faces from all images of both datasets;
2) Data Augmentation: we performed several image manipulations, such as rescaling and rotation in order to increase the diversity and quantity of our training set;
3) Classification Model: we apply transfer learning to a CNN architecture pre-trained with face images and also add fully connected layers specifically trained with neonatal face images, enabling pain assessment.

1) Facial Detection: We applied the state-of-the-art Retina Face [20] algorithm in all images of the UNIFESP and iCOPE datasets to extract the face from each image. It is a single-stage pixel-wise face localisation method that employs a multi-task learning strategy to simultaneously predict face score, face box, five facial landmarks, and 3D position and correspondence of each facial pixel. The results obtained in the original experiments by the authors outperformed existing methods and achieved average precision equal to 91.4%. Accordingly, all faces from UNIFESP and iCOPE datasets were detected.

2) Data Augmentation: It is well known that in order to successfully train a CNN, a significant amount of images is required. Therefore, for each face image of our dataset, we generated a total of 20 augmented images. Using Tensor Flow [21], we applied the following manipulations: rotation angle (30º), rescaling (0.15), horizontal (0.2) and vertical (0.2) offsetting, brightness (0.5 - 1.1), zoom (0.7 - 1.5) and horizontal flip.

3) Classification Model: Following results presented by Zamzmi et al. [14], we chose the VGG-Face architecture with 16 layers, originally proposed and implemented by Parkhi [15], as our classification model. Since we are dealing with a small dataset it is common to use the transfer-learning method, where we take advantage of an already pre-trained VGG-Face model, adding a fully-connected classifier on top, specifically trained with neonatal face images captured before and after the painful clinical procedure to classify the facial expression into two classes "pain" or "no pain", without pain level detection. We used Tensorflow [21] for training and testing our proposed CNN.

To find the best architecture, we performed a random search with parameters ranging from 50 to 2048 neurons and 1 to 3 fully connected layers. Using both datasets, the result suggested 2 fully connected layers with 512 neurons each. We used the Categorical Cross-Entropy with L1 regularisation penalty as the loss function and the RMSprop (Root Mean Square Propagation) [22] as our gradient descent optimisation algorithm. Experimentally learning rate ($\eta_{pre-ft} = 1e-4$ and $\eta_{during-ft} = 1e-6$), dropout (50%) and weight regularisation ($l_1 = 5e-4$) have also been selected, preventing overfitting.

It is noteworthy that all images must match the VGG-Face input size of 224 x 224 x 3. Also, Parkhi [15] changed RGB channels to BGR and centralised values of each channel on zero. An example is shown in Figure 1.

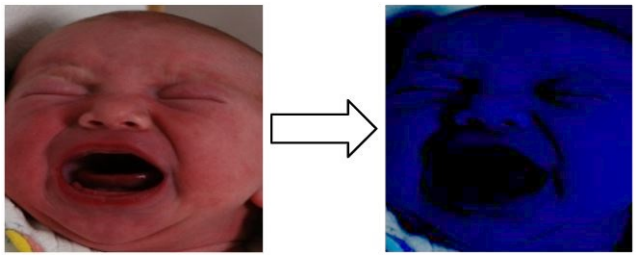


Fig. 1. VGG-Face pre-processing on an iCOPE image.

The training process was carried out to obtain three classification models: (1) trained with the UNIFESP dataset only, (2) trained with the iCOPE dataset only, and (3) trained with the UNIFESP + iCOPE dataset. However, all models were tested on both datasets (UNIFESP + iCOPE). Prior to fine-tuning, only weights of fully connected layers were adjusted.

Fine-tuning started when the result had not improved after 5 epochs, leading to weight adjustment of Groups 4 and 5 of the convolutional layers as well. We performed training with a batch size of 16. Each model was evaluated with the Hold-Out Cross-Validation technique (10 repetitions), randomly choosing independent test sets (20% of original datasets) for each repetition. The best model was optimised using the TensorFlow Lite Dynamic Range Quantisation function. The embedded model was quantised from floating-point weights (*float32*) to integers (*int8*) and during the inference process, weights are converted back from integers to floating points and cached in memory to reduce latency.

C. Mobile Application

We developed a mobile application capable of detecting a neonate's face and classifying it as "pain" or "no pain" with low latency and offline. Also, our software registers the performed analysis and makes them available for query, allowing metadata analysis and the collection of new face images. The classification model previously described had been optimised for mobile environment, considering the mobile devices' storage and processing specification. The application was developed for Android OS using the Android Studio IDE.

On the app's home screen, illustrated in Figure 2a, three buttons activate different functionalities of the application: instructions, camera, and history.

In Instructions, the user will find a quick guide for the app operation containing texts and images explaining the other two functionalities (camera and history).

Selecting the Camera button activates the app's main functionality: the neonatal facial expression classifier. Initially, in this mode, the user is asked to type the neonate's name to register the analysis that is going to be done. Then, the screen will show the back-camera's view, but with an oval shape on its center, which should be used to guide the user when positioning the neonate's face on the screen.

The real-time analysis starts right after the name insertion (Figure 2b). The camera automatically captures a picture, which is then processed by the face detector algorithm. In the mobile environment, the face detection is carried out by the Face Detection API from the Firebase's ML Kit, a Google platform for mobile and web applications development. If that algorithm could not detect a face in the captured image, the oval shape assumes a red colour (Figure 2c), indicating that the neonate's face positioning had not been done properly, and the app automatically captures another picture. In contrast, if a face was successfully detected, the oval shape is presented in green. The picture's region corresponding to the detected face is cropped out and inputted in the classification model, which will determine if the facial expression indicates the presence of pain or not. These screens are illustrated in Figures 2e and 2d. At the end of the classification model analysis, the user will see the image's classification, the result's confidence score, and the processing time. Then, the application automatically captures another picture, and the face detection and classification process are repeated. In order to interrupt this cycle and conclude the real-time analysis, the user needs to exit the Camera mode and get back to the app's home screen. To execute another analysis, the user must select the Camera button again.

At last, by selecting the History button at the home screen, the application will present a list containing the performed analyses register and a search tool, which can be used to find a specific neonate. This screen is shown in Figure 2f. The software always saves the first picture associated with the state change ("pain" or "no pain") of the subject's face. Therefore, in the Camera mode, after the analysis start, the first face to be detected will be saved, independently of the results. The next facial image to be saved is going to be the first one to receive a different classification from the last one. For instance, if in the first moment the neonate's face expression was classified as "no pain" (the first image to receive this denomination was saved), after detecting a state change ("pain"), the first face image will be saved. From this moment, the next face image to receive the "no pain" classification will be saved as well, and the cycle goes on until the user ends the analysis. This algorithm allows a highly scalable solution for depicting hundreds and thousands of neonatal face images.

Saved face images are associated with the neonate's name, the classification given, the reliability, and the date and time of the image capture, resulting in an analysis record. All records are grouped by the neonate's name, composing a profile for each patient. To view a specific profile, the user may type the neonate's name into the search tool. By clicking on the desired profile, the user will be directed to a screen containing graphs about the total number of analyses, days, and hours when the neonate was assessed as "pain" or "no pain". Figure 2g shows this profile.

TABLE I
EVALUATION METRICS RESULTS FOR EACH MODEL.

Metric	**UNIFESP**	**iCOPE**	**Both**
Accuracy	$72\% \pm 3\%$	$83\% \pm 2\%$	$89\% \pm 4\%$
Precision	0.85 ± 0.04	0.80 ± 0.03	0.86 ± 0.05
Sensibility	0.62 ± 0.06	0.93 ± 0.04	0.92 ± 0.05
F1	0.72 ± 0.04	0.86 ± 0.02	0.89 ± 0.04
AUC	0.74 ± 0.03	0.82 ± 0.03	0.86 ± 0.05

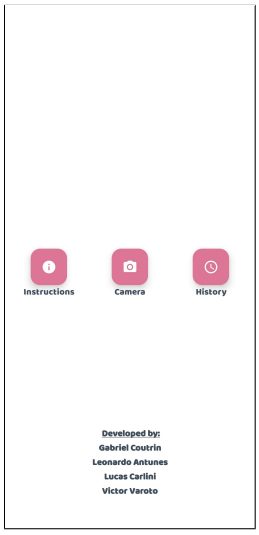
III. RESULTS AND DISCUSSION

In this section, we describe the results of our framework. Firstly, we present the evaluation of our classification model and, then, the results of model interpretability.

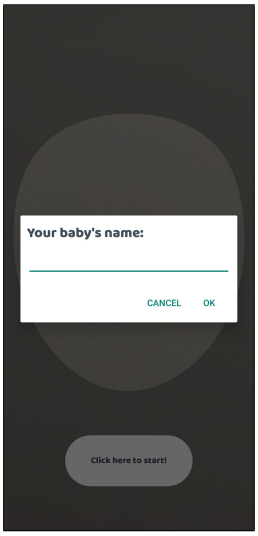
A. Evaluation Metrics

Table I shows results obtained for each classification model, as described at the end of Section II-B3.

We can see that, even though the UNIFESP model achieved 72% mean accuracy, the sensibility was relatively low (< 0.7). That means that the model has incorrectly classified a considerable amount of images as "no pain" (false negatives). This result also compromised the F1 Score and AUC. These metrics are equally important due to the serious consequences that under- or over-treatment of pain may cause to the neonate.

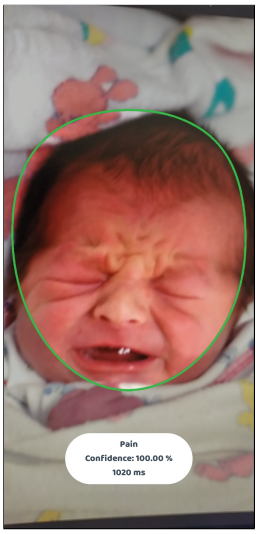


(a) Mobile application home screen.

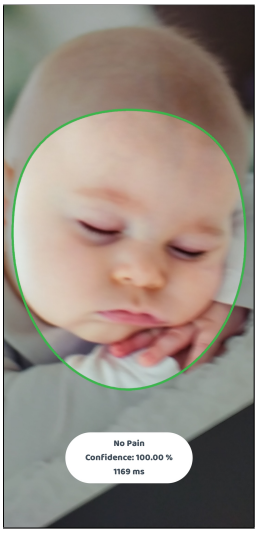


(b) Input neonate's name.

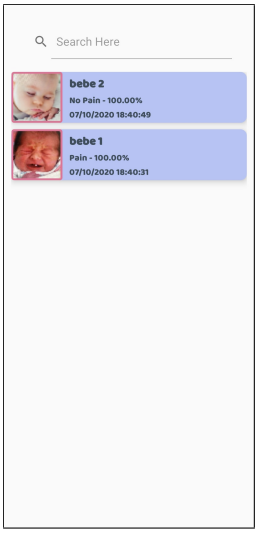
(c) Failed face detection.



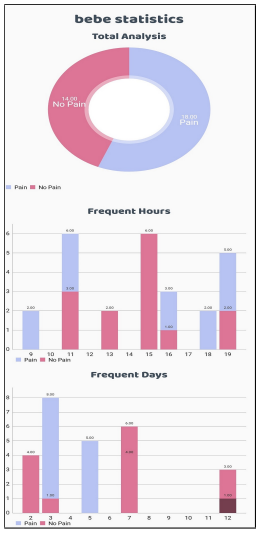
(d) Neonate with "pain".



(e) Neonate with "no pain".



(f) History.



(g) Individual analyses.

Fig. 2. Mobile application screens.

Afterward, we can observe that the iCOPE model presented substantial gain in sensibility, leading to better F1 Score and AUC when comparing to the UNIFESP model. Overall, the iCOPE model obtained a better performance than the UNIFESP model. Our hypotheses are that this better performance is due to the higher resolution of the iCOPE dataset or, perhaps, by the fact that all neonates of the iCOPE dataset are Caucasians. Further investigations are still necessary. Lastly, we can observe that the model trained with both datasets (UNIFESP + iCOPE) showed a better performance than both models analysed previously with 89% accuracy, 0.89 F1 Score, and 0.86 AUC. Specifically, the model obtained presented better metrics when comparing to the UNIFESP model, although it showed a slightly better performance than iCOPE model. We believe that these results are suited for neonatal pain assessment since commonly used clinical pain scales achieved similar results, such as the NFCS [10] that achieved reliability of 0.86.

We applied Analysis of Variance (ANOVA) [23] to verify the statistical difference between all models. The results suggest that UNIFESP + iCOPE model is the one that achieved better performance and differences between evaluation metrics are statistically significant at the $p = 0.05$ level.

Finally, we carried out a new training using UNIFESP + iCOPE dataset to better maximise evaluation metrics. This model achieved 93.07% accuracy, 0.9431 F1 Score, and 0.9254 AUC. As noted in Section II-C, we performed a mobile environment optimisation using the Dynamic Range Quantisation function available on TensorFlow [21]. The model had a reduction in memory space cost from $160Mb$ to $26Mb$ and

showed similar performance to the original. These results are shown in Table II.

TABLE II
FINAL CLASSIFICATION MODEL. EVALUATION METRICS PRIOR TO AND AFTER MOBILE ENVIRONMENT OPTIMISATION.

Model	Accuracy	Precision	Sensibility	F1	AUC
Prior	93.07%	0.9355	0.9508	0.9431	0.9254
After	93.07%	0.9107	0.9623	0.9358	0.9291

B. Model Interpretability

In order to better understand the relationship between our classification model prediction in terms of its features, we applied an explainable AI method, named Integrated Gradients [24]. The main idea of this technique is to accumulate pixel (image feature) local gradients and attribute its importance as a score for how much it adds or subtracts to the model's overall output class probability. To do so, we interpolate (α) images between a baseline image (e.g. black image with pixels equal to zero) and the input image. Then, we compute gradients to each interpolated image and we accumulate these local gradients, approximating the integral between the baseline and input image. Based on these accumulated local gradients, we have an attribution mask highlighting pixels relevant to the classification.

Figures 3 and 4 show some examples of Integrated Gradients applied to the test set of both datasets. Observing Figure 3, we can see that the most highlighted facial features are the forehead, upper contour of the nose, and nasolabial groove. Specifically to images classified as "pain", Integrated Gradients highlighted the mouth with tongue protrusion. Interestingly, however, is the fact that Figure 3b shows a neonate classified as "no pain" with 59.89% probability and with the mouth open. Finally, it is noteworthy that these images did not present gradients highlighting secondary artifacts of the image, such as the blanket, but were concentrated on the face of the neonate itself.

Figure 4 shows results on the UNIFESP dataset, highlighting similar facial features. Interestingly about Figure 4c, it shows a neonate with deep eye contours, and this feature was highlighted as well. However, as shown in Figures 4b and 4d, Integrated Gradients showed high dispersion in these images, highlighting, besides the face, other artifacts of it. As questioned before, the higher resolution of the iCOPE dataset may preserve better information for pain classification.

Analysing these results, we believe that the nasolabial groove, as well as the open mouth and protrusion of the tongue, maybe the most discriminating facial regions to pain assessment. Moreover, it is interesting that these features are deemed clinically relevant and agree with the visual perception of adults when assessing pain, whether they are health professionals or not [10], [25].

IV. MOBILE APPLICATION FEASIBILITY

Recognition of neonatal pain is a challenge for health professionals. The Mobile Convolutional Neural Network proposed is revolutionary in the sense that facial recognition of neonatal pain features will be independent of individual variation. The "human factor" in pain recognition may lead to wide variation and it depends on previous pain experience, affective mood and workload. In this context, the robustness of the proposed app may be one of the important features of its application in clinical practice.

Translational research need to be done in order to assess the accuracy of the app for different neonates and clinical situations in order to assess if the performance of the neural network is less prone to to subjective variables that modify pain assessment than the human performance.

V. CONCLUSION

This paper presents a novel computational framework implemented in a mobile environment that uses AI techniques to automatically identify the facial expression of pain in neonates, presenting feasibility in real clinical situations and practical for everyday use. Our findings showed promising results to correctly identify the facial expression of pain in neonates with high accuracy and generalisation capability.

Moreover, to the best of our knowledge, this is the first work to apply explainable AI techniques in neonatal pain classification. Our methodology presented novel results highlighting as well sound facial regions that agree with pain scales used by neonatologists and with the visual perception of adults when assessing pain in neonates, whether they are health professionals or not.

As future work, we intend to carry out practical tests of our mobile application in the Neonatal Intensive Care Unit of the São Paulo Hospital. We believe that these practical tests are needed to identify limitations of the proposed solution when dealing with difficulties of real situations, such as the presence of artifacts on the face of the newborn and physiological signal measuring instruments. We also intend to apply different explainable AI methods to better understand facial regions that might be relevant to pain assessment and use all the depicted face images to enlarge the face image samples and train our computational model. Finally, we are considering to evaluate more recent CNNs architectures, such as the Neonatal-CNN [16], the InceptionResNet [26], the DenseNet [27], the MobileNet [28], and ResNet [29], [30].

ACKNOWLEDGMENT

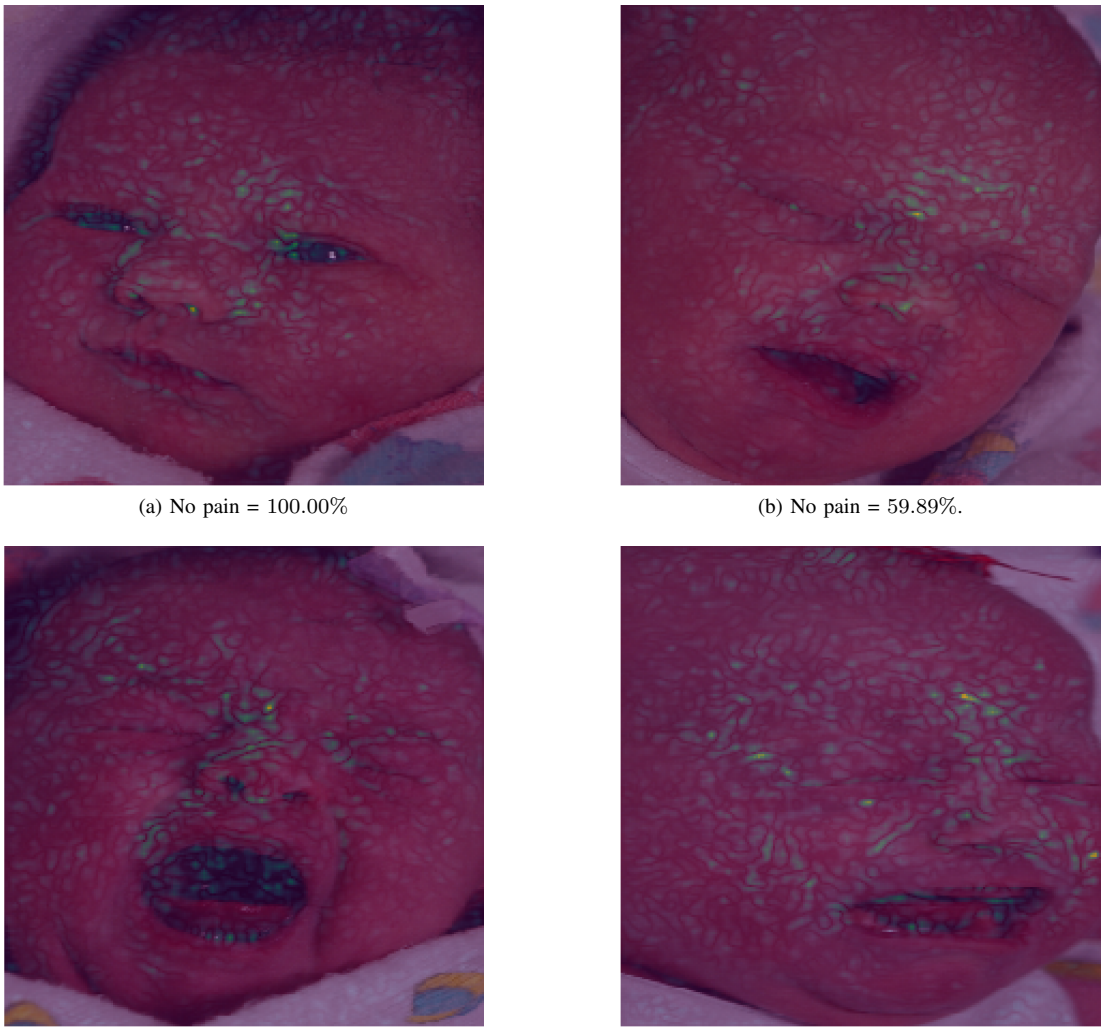
The authors would like to thank the financial support provided by the Brazilian funding agencies FAPESP (2018/13076-9), CNPq (401059/2019-7) and CAPES.

REFERENCES

[1] R. Guinsburg, "Avaliação e tratamento da dor no recém-nascido," *J Pediatr (Rio J)*, vol. 75, no. 3, pp. 149–60, 1999.

[2] R. E. Grunau, "Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity," *Rambam Maimonides medical journal*, vol. 4, no. 4, 2013.

[3] T. M. Heiderich, A. T. F. S. Leslie, and R. Guinsburg, "Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements," *Acta Paediatrica*, vol. 104, no. 2, pp. e63–e69, 2015.



(a) No pain = 100.00%

(b) No pain = 59.89%.

(c) Pain = 100.00%

(d) Pain = 99.99%

Fig. 3. Some examples of Integrated Gradients applied on iCOPE dataset and its corresponding classification probabilities.

[4] B. Golianu, E. J. Krane, K. S. Galloway, and M. Yaster, "Pediatric acute pain management," *Pediatric Clinics of North America*, vol. 47, no. 3, pp. 559–587, 2000.

[5] R. V. Grunau and K. D. Craig, "Pain expression in neonates: facial action and cry," *Pain*, vol. 28, no. 3, pp. 395–410, 1987.

[6] K. J. Anand, P. R. Hickey *et al.*, "Pain and its effects in the human neonate and fetus," *N Engl j Med*, vol. 317, no. 21, pp. 1321–1329, 1987.

[7] K. J. Anand and D. B. Carr, "The neuroanatomy, neurophysiology, and neurochemistry of pain, stress, and analgesia in newborns and children," *Pediatric Clinics of North America*, vol. 36, no. 4, pp. 795–822, 1989.

[8] R. C. Balda and R. Guinsburg, "Avaliação da dor no período neonatal," *Diagnóstico e tratamento em neonatologia. São Paulo: Atheneu*, pp. 577–85, 2004.

[9] F. A. M. Neves and D. A. M. Corrêa, "Dor em recém-nascidos: a percepção da equipe de saúde," *Ciência, Cuidado e Saúde*, vol. 7, no. 4, pp. 461–467, 2008.

[10] R. V. Grunau, C. C. Johnston, and K. D. Craig, "Neonatal facial and cry responses to invasive and non-invasive procedures," *Pain*, vol. 42, no. 3, pp. 295–305, 1990.

[11] G. F. Teruel, T. M. Heiderich, R. Guinsburg, and C. E. Thomaz, "Analysis and recognition of pain in 2d face images of full term and healthy newborns," *Proceedings of the XV Encontro Nacional de Inteligencia Artificial, ENIAC 2018*, pp. 228–239, 2018.

[12] C. E. Thomaz, J. P. Boardman, S. Counsell, D. L. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert, "A multivariate statistical analysis of the developing human brain in preterm infants," *Image and Vision Computing*, vol. 25, no. 6, pp. 981–994, 2007.

[13] I. Jolliffe and Springer-Verlag, *Principal Component Analysis*, ser. Springer Series in Statistics. Springer, 2002. [Online]. Available: https://books.google.com.br/books?id=_olByCrhjwIC

[14] G. Zamzmi, D. Goldgof, R. Kasturi, and Y. Sun, "Neonatal pain expression recognition using transfer learning," *arXiv preprint arXiv:1807.01631*, 2018.

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

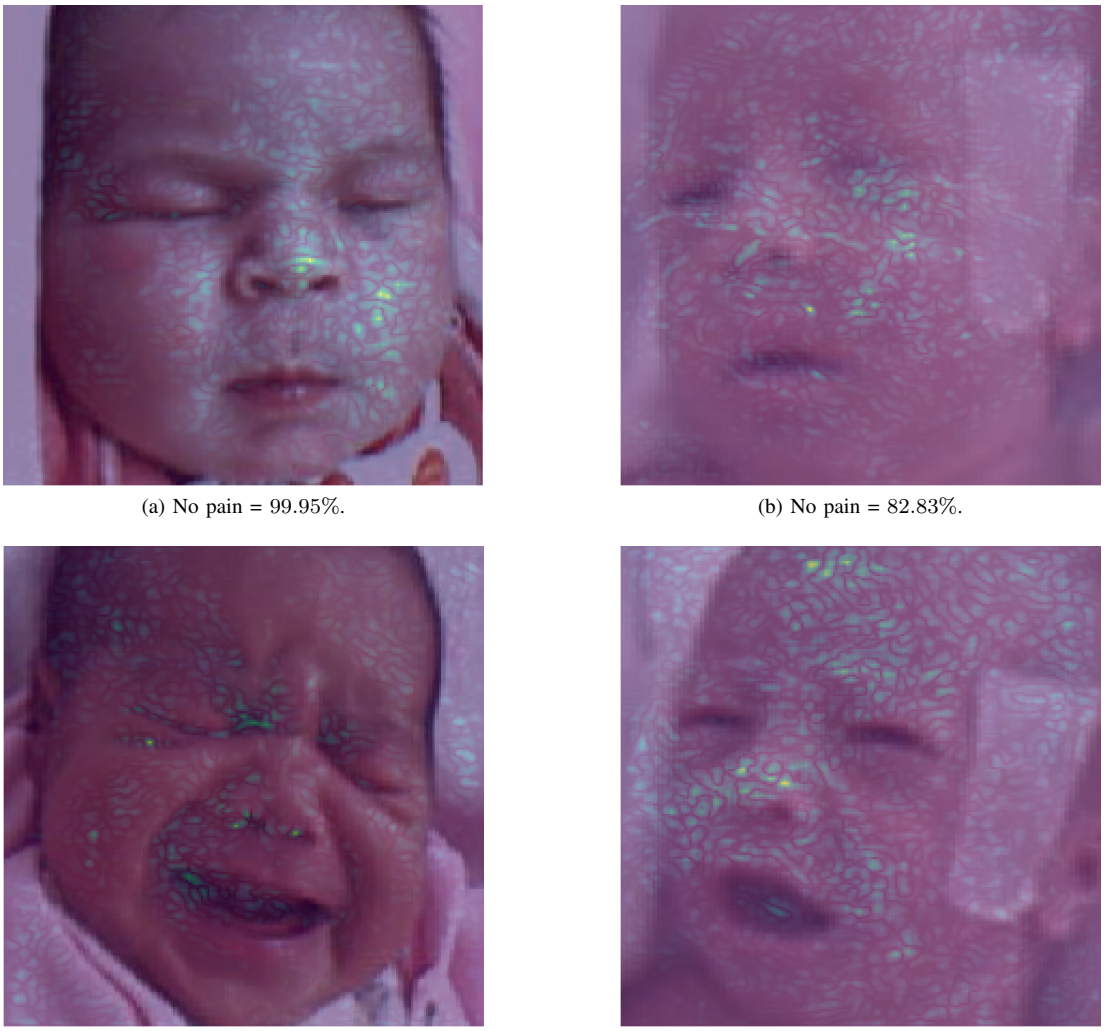
[16] G. Zamzmi, R. Paul, D. Goldgof, R. Kasturi, and Y. Sun, "Pain assessment from facial expression: Neonatal convolutional neural network (n-cnn)," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.

[17] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "Svm classification of neonatal facial images of pain," *International Workshop on Fuzzy Logic and Applications*, pp. 121–128, 2005.

[18] ——, "Machine recognition and representation of neonatal facial displays of acute pain," *Artificial intelligence in medicine*, vol. 36, no. 3, pp. 211–222, 2006.

[19] S. Brahnam, C.-F. Chuang, R. S. Sexton, and F. Y. Shih, "Machine assessment of neonatal facial expressions of acute pain," *Decision Support Systems*, vol. 43, no. 4, pp. 1242–1254, 2007.

[20] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface:



(a) No pain = 99.95%. (b) No pain = 82.83%.

(c) Pain = 100.00%. (d) Pain = 79.96%.

Fig. 4. Some examples of Integrated Gradients applied on UNIFESP dataset and its corresponding classification probabilities.

Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[21] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[22] G. Hinton, N. Srivastava, and K. Swersky, "Overview of mini-batch gradient descent," 2012, [Online; accessed 23-Julho-2020]. [Online]. Available: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

[23] R. Fisher, "On the'probable error'of a coeficient of correlation deduced from a small sample, metron i (1921)," *Reprinted in: Contributions to Mathematical Statistics, Wiley, New York*, pp. 3–32, 1950.

[24] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[25] L. P. Carlini, J. C. A. Soares, G. V. T. Silva, T. M. Heideirich, R. C. X. Balda, M. C. M. Barros, R. Guinsburg, and C. E. Thomaz, "A visual perception framework to analyse neonatal pain in face images," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, A. Campilho, F. Karray, and Z. Wang, Eds., vol. 12131. Cham: Springer International Publishing, 2020, pp. 233–243.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[30] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.