2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Facial landmark identification and data preparation can significantly improve the extraction of newborns' facial features

Giulio Del Corso[1,*], Danila Germanese[1,*], Maria Antonietta Pascali[1], Serena Bardelli[2], Armando Cuttano[2,3], Fabrizia Festante[4], Andrea Guzzetta[4,5], Lucia Rocchitelli[4,6], and Sara Colantonio[1]

[1] Institute of Information Science and Technologies, National Research Council of Italy (ISTI), Pisa, Italy
[2] Centro di Formazione e Simulazione Neonatale NINA, Azienda Ospedaliero Universitaria Pisana, Pisa
[3] UO Neonatologia, Azienda Ospedaliero Universitaria Pisana, Pisa
[4] Department of Developmental Neuroscience, IRCCS Stella Maris Foundation, Pisa, Italy
[5] Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy
[6] Department of Neuroscience, Psychology, Drug Research and Child Health NEUROFARBA, University of Florence, Florence, Italy

***Abstract*— Automatic extraction of facial feature can provide valuable information on the health of newborns. However, determining an optimal facial features extraction strategy, especially for preterm infants, is a challenging task due to significant differences in facial morphology and frequent pose changes.**

In this work, we collected video data from 10 newborns (8 preterm, 2 at term, ≤ 4 weeks post term equivalent age), obtaining a novel dataset of over $41,000$ labeled frames (*Open Mouth*, *Closed Mouth*, *Tongue Protrusion*). On the collected images, we applied a strong data preparation procedure (including mouth localization, cropping, and reorientation with models trained on adults), an adaptive image normalization strategy, and a proper data augmentation scheme. Thus, we trained a convolutional classifier with a large number of trainable parameters (i.e., $\sim$1.2 million), coupled with multiple criteria to avoid overspecialization and consequent loss of generalization capability.

This approach allows for highly reliable results (accuracy, precision, and recall over 92% on unseen data) and generalizes well to newborns with significantly different characteristics, even without including time-dependent information in the analysis. Therefore, these results prove that proper data preparation can narrow the gap between the classification of neonatal and adult facial features, allowing the integration of methods originally developed for adults into the complex setting of preterm infant analysis.

I. INTRODUCTION

Children's faces contain a wealth of valuable information regarding their health. Indeed, certain pathological conditions alter the expression or appearance of children's faces due to physiological or behavioral reactions [1], [2]. Contactless approaches, such as computer vision methods, may detect and analyze the most relevant facial features, thus providing clinicians (or parents, teachers, caregivers, etc.) with unobtrusive and objective information about children's health status [3]. The developed methods range from classification of infants' facial expression configuration [4], assessment of motor disorders [5], or even video-based behavior analysis for autism diagnosis [6].

However, the challenge of this task increases as the age of the newborns studied decreases, due to significant changes in the morphology of the face compared to an adult face and increased difficulty in data acquisition due to random changes in their facial pattern and pose [7], [8]. Consequently, most of the available datasets are focused on older infants (e.g., $\geq$6 months [9], [10], $\geq$2 years [11], [12]) or include only images instead of videos [13]. Therefore, research problems that focus on early newborns (i.e., less than 4 months from birth), especially preterm and late-preterm infants, are particularly challenging.

Among the research questions related to early newborns, one open issue concerns neonatal imitation (NI), meant as the existence of a primitive ability of infants to mirror the actions of others [14], [15]. The question of whether imitation is present from birth is still open and debated in the scientific community, though being of great importance [16], as it can foster a more nuanced understanding of how imitation serves as a building block for later developmental outcomes. This is especially important for preterm infants, being at risk of impaired neurophysical development.

To support investigations into NI, a crucial task is to detect and assess specific Facial Action Units (FACs, [17]) in response to a matching stimulus. In this respect, the detection and tracking of facial landmarks is a basic step; although well-established tools can track facial gestures in adults [18], they are not really effective when applied to the recognition and tracking of facial gestures in newborns due to critical differences in morphology and pose dynamics [19].

In this work, we approach the problem by analyzing videos of newborns (8 preterm, 2 at term, ≤ 4 weeks post term equivalent age) performing different tasks (tongue protrusion, mouth opening, etc.) to classify open/closed mouths. The videos are analyzed at frame-level. First, we identified mouth landmarks and cropped the images around the mouth, then we applied a rigorous normalization procedure (mouth orientation, resizing, brightness, and contrast enhancement) to make the dataset homogeneous and guarantee better classification performance. Appropriate strategies based on resampling/data augmentation, and unequal class weights are implemented to compensate for inhomogeneous acquisition

979-8-3503-9494-8/24/$31.00

Fig. 1: Setting of data acquisition. The image shows a 3D render of the room showing the equipment used and its positioning in relation to the operator and the infant.

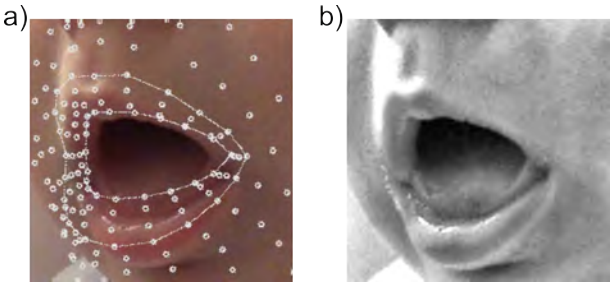

Fig. 2: Data preparation procedure: the original image is processed using Face Landmarker of Google MediaPipe Solutions to identify a rough contour of the mouth (a). This imprecise contour is used to crop/reorient the image. An adaptive brightness/contrast enhancement is applied to the final image (b).

lengths among subjects. A Convolutional Neural Network (CNN), coupled with an appropriate early stopping strategy, is trained using a ten-fold cross-validation to provide deeper insight into classification capability.

II. DATA ACQUISITION AND PREPARATION

A. Ethical approval

The study was approved by the Tuscany Region Pediatric Ethics Committee (123/2020, 12/2021) and conducted in accordance with the ethical principles of the Declaration of Helsinki. Families gave written informed consent before participating in the study.

B. Inclusion criteria

Preterm infants were recruited according to the following inclusion criteria: (1) born between 32 and 36 gestational weeks, admitted to the Neonatal Unit of the University Hospital of Pisa, Italy, (2) stable clinical conditions, and (3) none or minor brain abnormalities on ultrasound (transient flare, mild isolated ventricular dilation). Full-term infants were recruited from the Baby Nursery of the same hospital, according to the following inclusion criteria: (1) born $\geq$ 37 gestational weeks, and (2) absence of perinatal complications.

C. Data acquisition

The video recordings were conducted in a well-equipped room located close to the neonatal unit, but far enough away to allow for the absence of other newborns' cries. The entire recording session comprised three conditions: tongue protrusion, mouth opening, and a control disk. During each condition, an operator positioned in front of the newborn alternated between presenting dynamic stimuli and static face (or disk) periods at predefined time intervals and for a total duration of 3 minutes. The order of conditions was randomized across newborns. The entire recording session lasted 9 minutes. The room was equipped with two light points characterised by warm light, placed behind the neonatal station where the infant was placed as shown in Fig. 1. This choice was made to prevent the light from creating a point of attraction for the newborn and thus distracting from the execution of the task, while at the same time clearly illuminating the face of the operator performing the predefined movements. The newborn was placed on a bassinet equipped with a pillow to allow a comfortable and inclined position to improve the operator's visibility as show in Fig. 1. The main video camera used was a Canon Legria HFG70, a professional CMOS 4K 1/2.3 video camera, which, thanks to its advanced autofocus, 20x optical zoom and 5-axis stabilisation, was excellent for creating smooth, accurate and high-definition video. The camera was placed on a tripod and manually moved by a second operator to keep the lens on the baby's face as much as possible as show in Fig. 1. In order not to interfere with the performance of the task and to avoid distracting the newborn, a selected position was chosen to the side of the neonatal bassinet and behind the operator. A second camera was however set up on the other side of the neonatal bassinet to allow another more focused shot of the operator's face.

D. Mouth identification

The collected recordings are divided into frames, which in most cases include the whole face of the newborn and the surroundings, but sometimes the face of the newborn is covered. Therefore, as a pre-processing step, we want to identify when the newborn's mouth is visible and then crop the image to include only relevant information to correctly classify the open/closed mouth.

Although face detection and characterization in infants is a challenging task in computer vision [19], methods developed for adult facial landmark estimation are sufficiently accurate to produce a reasonable crop around the mouth of newborns. Our identification of the mouth exploits the Face Landmarker of Google MediaPipe Solutions [20], which first uses the Blazeface model [21] to detect faces, then uses a second model to locate mouth landmarks on the detected faces. These landmarks are used to define the mouth length and rotation to the respect of the image horizontal axis. Frames that do not contain mouth landmarks are removed from the dataset.

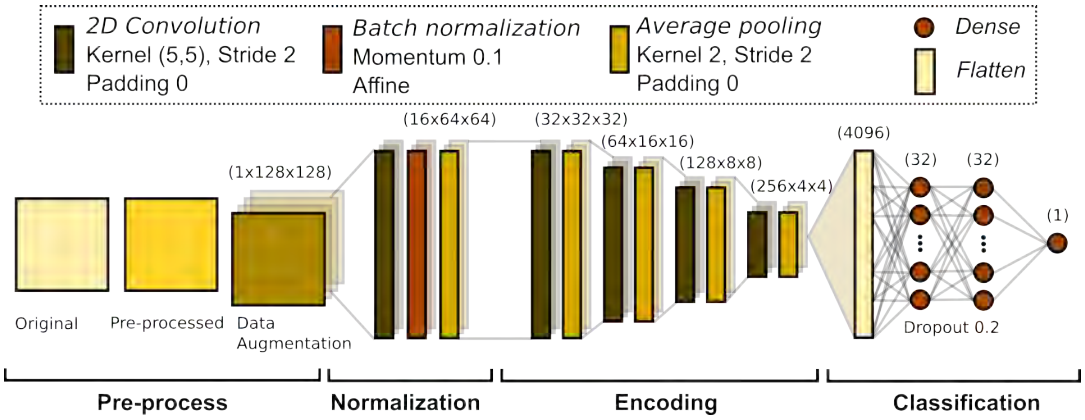

Fig. 3: Scheme of the Convolutional Artificial Neural Network classifier.

In our experimentation, the infant generally moves and rotates the head during the acquisition; then, in order to have a collection of mouth more homogeneous, we extracted the rotated bounding box of the mouth in each frame and stored it. In other words, for each frame of a video sequence, we saved a crop of the image representing the infant's mouth, and oriented in the direction {left mouth corner - right mouth corner}, see Fig. 2a).

E. Image pre-processing

Cropped images were pre-processed to enhance their textural information while preserving their quality and facilitating the subsequent analysis, as reported in Fig. 2b). Preprocessing steps were as follows:

- *Image resizing:* All images were reduced to a predetermined size (128x128 pixels), to lighten the computational load of downstream analysis.
- *Brightness enhancement:* The brightness of each image was first assessed according to [22]. In the case of low values, the brightness was increased by a brightness factor.
- *Grayscale conversion:* All images were converted from RGB to grayscale images to reduce the number of input channels of the model and drastically reduce the trainable parameters.
- *Contrast enhancement:* A Contrast Limited Adaptive Histogram Equalization (CLAHE) was implemented, to improve visual discriminability and detail rendition without increasing the signal-to-noise ratio [23]. CLAHE, a variant of traditional histogram equalization, operates by partitioning an image into smaller tiles and redistributing pixel intensities within each tile to achieve a uniform histogram. By constraining the amplification of local contrast through adaptive clipping, CLAHE mitigates the risk of over-enhancement and preserves the natural appearance of images. Consequently, this technique may improve the analysis of complex textural features, particularly in those images characterized by non-uniform illumination or low contrast gradients.

F. Data labelling

Images in which the mouth is totally covered (usually by the newborn's hand) or in those in which the automatic detection of mouth landmarks fails are excluded from the dataset. The remaining images are then consensus labeled by 3 experts, classifying each image as *Closed Mouth*, *Open Mouth*, or *Tongue Protrusion*. For this preliminary work, the last two classes are collapsed to a single one, still called *Open Mouth*. Images that do not have a consensus prediction are excluded from the dataset.

III. MODEL DESCRIPTION AND VALIDATION SCHEME

A. Model description

We applied a deep convolutional classification model to identify open/closed mouths at frame level. As reported in Fig. 3, the architecture includes a convolutional layer followed by a batch normalization one to reduce internal covariate shift. This layer, combined with a high batch size ($n = 40$) and a reduced momentum (0.1), allows to maintain a high learning rate and thus drastically speed up the training phase [24]. Normalization is followed by average pooling (kernel size 2) to reduce the image size and reduce the degrees of freedom of the network.

The normalization block is followed by four encoding blocks (made up of a convolutional layer (kernel size 5, stride 1, padding 2), a ReLU activation function, and an average pooling (kernel size 2) that transforms the original 128×128 image into a 4×4 one. These images are flattened to obtain a one-dimensional tensor of 4096 elements.

The classification block of the network consists of two dense layers of 32 nodes each, providing a single output. Each of these layers is preceded by a dropout layer (Bernoulli distribution, $p = 0.2$) to further reduce early overfitting phenomena. Instead of applying a sigmoid transform, we take advantage of the log-sum-exp trick of a Binary Cross Entropy with Logit loss to increase numerical stability.

The final model, optimized using an ADAptive MOment estimation method (ADAM), has 1,220,098 trainable parameters.

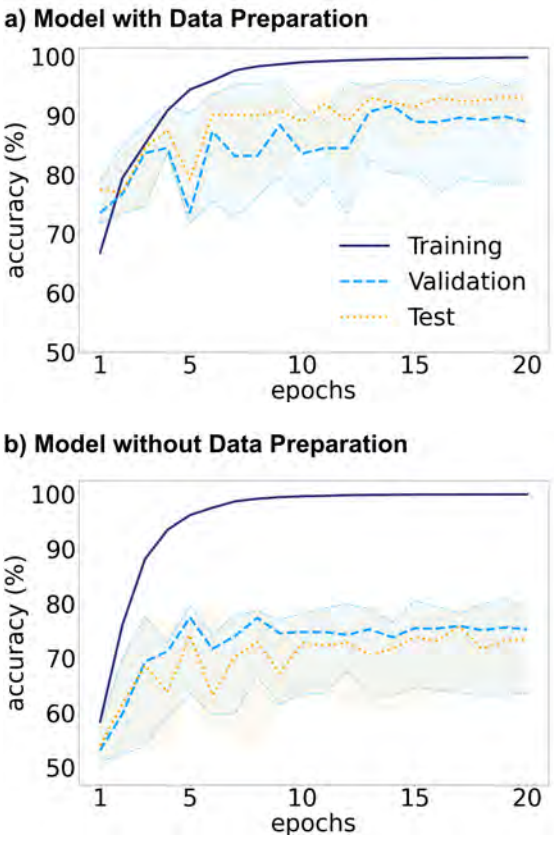

Fig. 4: Model convergence on training, validation and test sets for the model with (a) and without (b) data preparation. Values are reported as median and interquartile range among the 10 folds.

B. Stratification, Data Augmentation, and Balancing of Newborns

The labeled images are obtained from video acquisition. Therefore, there is a very high similarity between consecutive frames, which, if not addressed correctly, can produce positively biased results. To overcome this limitation, we stratified the dataset at subject-level. Images belonging to a given newborn are never split between different sets (training/validation/test), ensuring an unbiased evaluation of model performance.

Providing slightly different consecutive images of the same newborn can be perceived as data augmentation, but this can reduce the ability to generalize to unseen data by over-specializing the model to long series of similar images.

Therefore, we decided to add random perturbations to the images to prevent model overfitting (data augmentation). In particular, for each image, a value $u \in \mathcal{U}[0.7, 1]$ is sampled from an uniform distribution. A sub-image with area $u \cdot$ *Original Area*, same aspect ratio and center (c_x, c_y) is extracted from the original image. The center is randomly sampled to guarantee that the sub-image is completely contained in the original image. A bilinear interpolation is applied to the sub-image in order to map it to the resolution of the original image. We also applied a standard horizontal flip (vertical axis of the simmetry) with a probability of 50% to further balance image similarity.

The dataset is therefore expanded by a factor of 2 (horizontal flip), and the same image is never given to the model twice, as the random online sampling of subimages always adds a certain amount of variability to the training data. In addition, the batch operates on a shuffled version of the entire image dataset to avoid providing similar images in the same backpropagation step.

This dataset is characterized by strong heterogeneity among labeled frames for each newborn (from a minimum of 537 to a maximum of 8181). This can potentially affect model performance, as newborns with a higher number of available images will have a greater effect during training than those with fewer images. Therefore, we applied a weighted loss based on the number of available images for each newborn to artificially inflate the impact on model training of newborns with short video acquisition.

TABLE I: Model performance in validation and test set. Results are reported as median and interquartile range for both models (with and without data preparation).

	Accuracy	Precision	Recall
Validation (Data Preparation)	**92.0%** [83.9, 97.3]	**91.7%** [84.5, 96.9]	**96.9%** [91.3, 97.5]
Test (Data Preparation)	**92.2%** [85.3, 96.0]	**92.3%** [83.7, 96.6]	**95.4%** [92.0, 97.0]
Validation	75.2% [64.1, 78.3]	70.9% [60.4, 77.9]	76.0% [68.3, 87.4]
Test	72.5% [61.5, 75.6]	65.0% [59.1, 75.4]	77.9% [63.1, 83.4]

C. Training and validation scheme

To reduce the risk of early over-specialization of the model at local minima, we define a triangular adaptive learning rate that ranges linearly from an initial value of 0.01 to a maximum value of 0.05 (warm-up steps). Next, the value is linearly reduced from the maximum to 0.001 for 15 epochs (to mitigate exploding gradient problems).

In order to take advantage of the structure of the database (reduced number of newborns (10), but each with a high number of labeled images (3278 median number of available slices per newborn)), we implemented a leave-one-out (LOO) cross-validation strategy at subject-level, which proves effective in determining model performances in biomedical settings [25]. Each LOO fold divides the data set into training (80%), validation (10%), and test (10%) sets. Accuracy on the validation set is used as an early stopping criterion (the best model is defined as the one that maximizes validation accuracy after the warm-up steps), performance is then evaluated on the test set. This procedure, even if costly in terms of simulation, allows to obtain a low-biased estimate on generalization capability of the model.

IV. RESULTS

A. Study population

Eight preterm newborns (7 male, 1 female) and 2 full-term newborns (1 male, 1 female) were involved in this

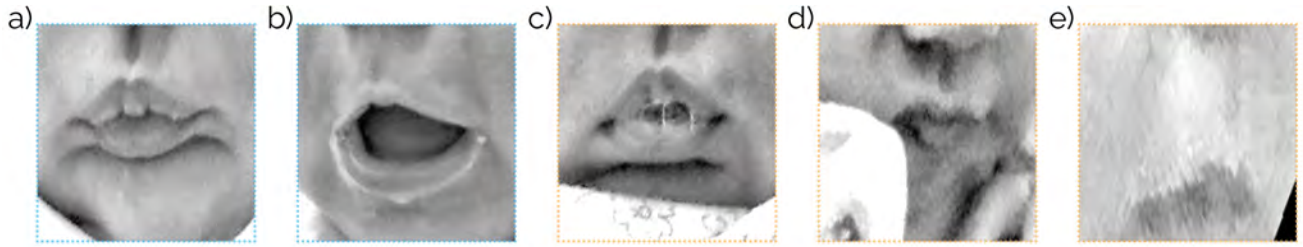

Fig. 5: Examples of image post Data preparation. a,b) show frames with good model prediction, while c,d,e) depicts a recently fed newborn, a mouth partially covered by the operator, and motion artifacts.

study. The preterm infants' gestational ages at birth ranged between 33+4 and 36+0 gestational weeks, and between 35+4 and 43+2 weeks at the time of their participation in the video recording session. The two full-term newborns were born at 39+1 and 38+4 gestational weeks, respectively; and participated in the study at 39+6 and 42+1 weeks, respectively. All infants were assessed before reaching 4 weeks post-term equivalent age.

B. Model convergence

The full evaluation of the model involves training 10 models (on an average of original 26,224 images before data augmentation) with subsequent evaluation of performance on the validation set, which is used to define an early stopping criterion, and subsequent unbiased estimation on the test set.

Fig. 4 a) shows the median accuracies for training, validation, and test across epochs (and their interquartile ranges over the folds). In particular, the early stopping criterion is met after a median of 13 epochs (IQR [12, 15]) and is never met before the first epoch or after the 18th epoch.

Indeed, after the 10th epoch (5 warm-up+5 decay epochs), the models stabilize their performance and reach a high accuracy in both validation and training. The comparable performance (no statistically significant difference) on these two sets proves the convergence of the model and ensures a reduced risk of overfitting.

C. Classification results

The median models performance on the validation and test sets are shown in Tab. I. Considering the high number of degrees of freedom of the trained networks (1,220,098 trainable parameters), the training metrics are always around 100%. The trained model as an overall accuracy of 92.2% (on unseen test data), with a very low number of false negative (recall/sensitivity of 95.4%), i.e. open mouth classified as closed ones.

It should be emphasized that in 40% of the dataset, the validation/test performance reaches an excellent accuracy of over 97/95%, while the median scores are lowered by a few newborns (20%) with reduced model performance.

D. Subject-level analysis

A focus on newborns with out-of-distribution model results can be very helpful for subsequent improvement of the classification model. In particular, two newborns had unsatisfactory performance. These infants prove to exhibit anomalous recording, with a noisy image quality (Fig. 5 e)and with most of the frames containing tools from the acquisition settings that partially cover the newborn's mouth (Fig. 5 d). Similarly, even when performance is good, images of newborns who have recently fed (resulting in milk bubbles and residual regurgitation during recording, see Fig. 5 c) exhibit poor automatic labeling.

E. Effect of data preparation

To evaluate the effect of data preparation (cropping, rotation, adaptive enhancement, and *ad hoc* augmentation), we applied the same ANN to the data without any kind of preprocessing and preparation. The convergence curve (Fig. 4 b) shows a similar behavior compared to the data that underwent the full data preparation, but reaches a reduced performance on the test data (accuracy 72.5%, precision 65%, recall 77.9%). These performances, which are stable after the 10^{th} epoch of training, prove the importance of proper data preparation to clean the unnecessary information from the images and maximize the network's ability to generalize to unseen data.

V. CONCLUSIONS AND FUTURE WORKS

In this work, we applied a convolutional neural network coupled with a strong preprocessing procedure (mouth identification with pre-trained face landmark model and image normalization with adaptive brightness and contrast enhancement) to classify open/closed mouths in newborns undergoing tongue protrusion tasks.

After the data preparation procedure, based on automatic cropping and reorientation of the frames using a method pre-trained on adult facial features, we applied a robust normalization strategy (adaptive brightness/contrast enhancement). This procedure, combined with a very high degree of data augmentation, allows us to obtain reliable results (accuracy, precision and recall over 92% on unseen test data). Furthermore, the models behaves coherent among most of the newborns (accuracy over 90% for 60% of the test data). On the contrary, providing the model with unprocessed images leads to a reduced ability to discriminate open/closed mouths due to the heterogeneous information content and the dissimilarities between the frames. This demonstrates the importance of robust data preparation and the potential use

of methods suitable for adult features as a preprocessing step in the classification of newborn facial features.

Newborns with reduced model performance show anomalies in the acquisition procedure (reduced resolution due to the small area of the camera including the mouth and classification noise induced by frames with milk regurgitation), which can be mitigated by slight changes in the experimental setup. In fact, recording for automatic analysis is complicated by the natural movement of the infant's face that the operator must follow, and obtaining an unobstructed camera line is complex. Therefore, the recording cameras should not be set at a fixed distance, as this setting may be effective for behavioral purposes but not for automated analysis, and should follow the infant's face to ensure that the newborn's mouth covers a certain percentage of the recorded image. In addition, recently fed newborns should be excluded from the training procedure and thus from the automatic classification.

This approach is a first step towards defining a model capable of correctly classifying mouth-related facial features of newborns/preterm infants from video sequences. Indeed, the two main limitations of this preliminary work are: the single frame-level analysis and the non-inclusion of the protruding tongue class due to their under-representation in the study sample. Therefore, future developments of this work will focus on protrusion using temporal information both before and after the frame under consideration (recurrent neural network/long-term short-term memory model [26], [27]), not only to improve the model's predictive capabilities, but also to define a reliability score useful for labeling the uncertain transitions between states that characterize this type of data. In addition, the database will be expanded to include more newborns (4-10), at least doubling the number of images available.

The uncertain class can also be derived from network training based on multiple labelings of the same video sequences (e.g., from different research centers), thus incorporating human classification variability [28]. The development and integration of these methods into an already pre-trained, downloadable package will allow the automation of this labeling process in the analysis of infant imitation processes, objectifying an otherwise very complex and potentially biased procedure.

ACKNOWLEDGEMENTS

This publication is partly based upon work from COST Action *GoodBrother—Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living* (CA19121), supported by COST (European Cooperation in Science and Technology).

This work was supported by the Italian Ministry of Health - Grant RC L1 (and the 5x1000 voluntary contributions). AG was partially supported by Horizon 2020 project *BornToGetThere* no. 848201.

REFERENCES

[1] Dhanya Lakshmi Narayanan, Prajnya Ranganath, Shagun Aggarwal, Ashwin Dalal, Shubha R Phadke, and Kaushik Mandal. Computer-aided facial analysis in diagnosing dysmorphic syndromes in indian children. *Indian Pediatrics*, 56:1017–1019, 2019.

[2] Christa Einspieler, Arend F Bos, Melissa E Libertus, and Peter B Marschik. The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction. *Frontiers in psychology*, 7:178796, 2016.

[3] Danila Germanese, Sara Colantonio, Marco Del Coco, Pierluigi Carcagnì, and Marco Leo. Computer vision tasks for ambient intelligence in children's health. *Information*, 14(10), 2023.

[4] Andreas Maroulis. Baby facereader au classification for infant facial expression configurations. *Measuring Behavior 2018*, 2018.

[5] Elisa Ferrer-Mallol, Clare Matthews, Madeline Stoodley, Alessandra Gaeta, Elinor George, Emily Reuben, Alex Johnson, and Elin Haf Davies. Patient-led development of digital endpoints and the use of computer vision analysis in assessment of motor function in rare diseases. *Frontiers in Pharmacology*, 13:916714, 2022.

[6] Abid Ali, Farhood F Negin, Francois F Bremond, and Susanne Thümmler. Video-based behavior understanding of children for objective diagnosis of autism. In *VISAPP 2022-17th International Conference on Computer Vision Theory and Applications*, 2022.

[7] Dana Kuefner, Viola Macchi Cassia, Marta Picozzi, and Emanuela Bricolo. Do all kids look alike? evidence for an other-age effect in adults. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4):811, 2008.

[8] Ian WR Bushnell. Mother's face recognition in newborn infants: Learning and memory. *Infant and Child Development: An International Journal of Research and Practice*, 10(1-2):67–74, 2001.

[9] Mercedes Torres Torres, Michel Valstar, Caroline Henry, Carole Ward, and Don Sharkey. Postnatal gestational age estimation of newborns using small sample deep learning. *Image and vision computing*, 83:87–99, 2019.

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

[11] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 5:127200, 2015.

[12] Zakia Hammal, Wen-Sheng Chu, Jeffrey F Cohn, Carrie Heike, and Matthew L Speltz. Automatic action unit detection in infants using convolutional neural network. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 216–221. IEEE, 2017.

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

[14] Andrew N Meltzoff and M Keith Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75–78, 1977.

[15] Andrew N Meltzoff and M Keith Moore. Newborn infants imitate adult facial gestures. *Child development*, pages 702–709, 1983.

[16] Jacqueline Davis, Jonathan Redshaw, Thomas Suddendorf, Mark Nielsen, Siobhan Kennedy-Costantini, Janine Oostenbroek, and Virginia Slaughter. Does neonatal imitation exist? insights from a meta-analysis of 336 effect sizes. *Perspectives on Psychological Science*, 16(6):1373–1397, 2021.

[17] Harriet Oster. Baby facs: Facial action coding system for infants and young children. *Unpublished monograph and coding manual. New York University*, 2006.

[18] Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and CKM Lee. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer Science Review*, 48:100545, 2023.

[19] Michael Wan, Shaotong Zhu, Lingfei Luan, Gulati Prateek, Xiaofei Huang, Rebecca Schwartz-Mette, Marie Hayes, Emily Zimmerman, and Sarah Ostadabbas. Infanface: Bridging the infant–adult domain gap in facial landmark estimation in the wild. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4486–4492. IEEE, 2022.

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[21] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

[22] BT Series. Studio encoding parameters of digital television for standard 4: 3 and wide-screen 16: 9 aspect ratios. *International Telecommunication Union, Radiocommunication Sector*, 2011.

[23] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[25] Giulio Del Corso, Danila Germanese, Claudia Caudai, Giada Anastasi, Paolo Belli, Alessia Formica, Alberto Nicolucci, Simone Palma, Maria Antonietta Pascali, Stefania Pieroni, et al. Adaptive machine learning approach for importance evaluation of multimodal breast cancer radiomic features. *Journal of Imaging Informatics in Medicine*, pages 1–10, 2024.

[26] Yuan Liao, Linxuan Zhang, and Chongdang Liu. Uncertainty prediction of remaining useful life using long short-term memory network based on bootstrap method. In *2018 ieee international conference on prognostics and health management (icphm)*, pages 1–8. IEEE, 2018.

[27] Qingxiong Tan, Mang Ye, Andy Jinhua Ma, Baoyao Yang, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and Pong C Yuen. Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4665–4679, 2020.

[28] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.