

Deep learning for medical image analysis: from diagnosis to treatment and follow up

Daniele Berardini¹, Alessandro Cacciatore², Mariachiara Di Cosmo¹, Maria Chiara Fiorentino¹, Giovanna Migliorelli³, Lucia Migliorelli¹, Francesca Pia Villani^{2,*}, Emanuele Frontoni⁴ and Sara Moccia^{5,6}

¹Department of Information Engineering, Università Politecnica delle Marche, Italy

²Department of Humanities, Università degli Studi di Macerata, Italy

³Department of Law, Università degli Studi di Macerata, Italy

⁴Department of Political Sciences, Communication and International Relations, Università degli Studi di Macerata, Italy

⁵The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy

⁶Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy

Abstract

Deep Learning (DL) has shown to be a powerful tool for medical image analysis, with applications ranging from diagnosis to treatment and followup. In this work, we present an overview of the field, with a focus on four specific applications. The first application describes a contrastive learning approach for semi-supervised classification of fetal standard planes in ultrasound images, to help medical professionals to identify anatomical landmarks and measure fetal growth. We then describe a system for the automatic detection of stenosis in coronary angiography images, which can aid in the diagnosis and treatment of cardiovascular diseases. In this application we also provide a focus on federated learning, using data from two different medical centres to improve accuracy and generalizability. A further application regards the development of a DL model for cancer segmentation in videoendoscopic frames of the larynx, to assist clinicians in cancer early detection and treatment. Finally, we show a DL-based system to automatically monitor preterm infants in neonatal intensive care units. The system makes use of images acquired with an RGB-D camera placed on top of the crib, to estimate the positions of anatomical regions of interest. In this section we also give a perspective on green artificial intelligence, to develop fairer technologies, and on the ethical aspects related to the use of DL in the actual clinical practice.

Keywords

Medical image analysis, Deep learning, Computer-assisted diagnosis, Context awareness

1. Introduction

During the last decades, deep learning (DL), and in particular convolutional neural networks (CNNs), have undergone an increasing role in medical image analysis to offer decision support and context awareness to clinicians, and today an extensive literature exists [1].

Current challenges in the field include the high intra- and inter-patient variability, the lack of standardized guidelines to perform image annotation, the difficulty to gather annotation from experts, and the paucity of publicly available datasets, when compared to other fields, such as natural image analysis. The reason behind these challenges can be seen in the intrinsic nature of the images, which are considered sensitive data, with relevant legal and ethical issues, and, to be annotated, require efforts from clinicians that are already overwhelmed by their daily activity. Furthermore, each medical imaging modality brings its own additional characteristics, such

as low signal-to-noise ratio or poor boundaries [1].

This contribution summarizes main research activities of our group in the field of medical image analysis, ranging from diagnosis to treatment and followup.

2. Fetal ultrasound imaging

Identifying standard planes during fetal ultrasound acquisition is the first step for fetal biometry measurement and organ evaluation. Today, this task is performed by clinicians, who move the probe across the mother's belly, searching for specific anatomical landmarks. Typically acquired planes include maternal cervix, fetal femur, thorax and brain. Brain standard planes can be further classified in trans-ventricular (TV), trans-thalamic (TT) and trans-cerebellar (TC). Even though an extensive literature on this theme already exists [2], the problem of relying on time-expensive data annotation has not been fully solved in this domain. To mitigate the issue, in other fields of medical image analysis, contrastive representation learning based on instance discrimination tasks has gained much attention to incorporate unlabelled data in the training phase and improve classification performance despite

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

✉ f.villani2@unimc.it (F. P. Villani)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



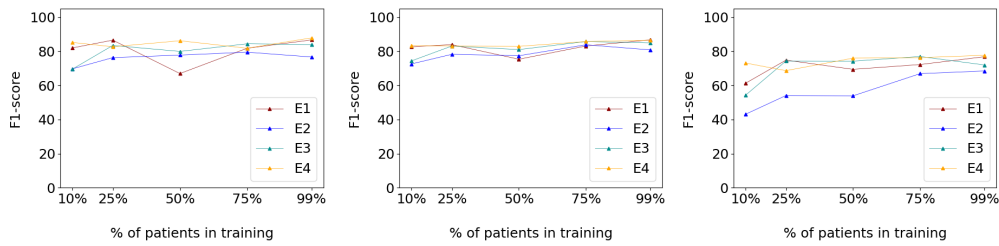


Figure 1: F1-score obtained with E1, E2, E3 and E4 applied to Z-brain and considering different percentage of training data. From left to right, results are shown for trans-cerebellar, trans-thalamic and trans-ventricular.

the limited amount of annotations. However, in the field of fetal US, the advantages of this approach are still questioned.

We study contrastive learning for semi-supervised classification conducting a fair comparison among:

- E1 pre-training on ImageNet followed by supervised finetuning: pretrained ResNet50 backbone using ImageNet weights, to assess the benefit of transfer learning using a large natural-image dataset.
- E2 self-supervised pre-training: ResNet50 backbone trained from scratch using simCLR [3], to evaluate the effectiveness of initializing backbone weights using contrastive learning.
- E3 self-supervised pre-training starting from ImageNet weights: E3 replicates E2 but training from scratch is replaced with pretraining using ImageNet weights, to assess whether combining transfer learning and contrastive learning may be beneficial.
- E4 End-to-end dual task architecture: this training includes a pretext-task based on contrastive learning and a classifier, both sharing the same backbone whose initial weights are those from ImageNet.

We investigate how the classification performance changes when considering different percentages of training patients, and datasets with high intra-class variability ($Z - multiclass$) and low intra-class variability ($Z - brain$). The analyses are performed on the most recent publicly-available dataset in the field¹. Classification performance is assessed quantitatively by means of F1-score to deal with class imbalance. ANOVA test is performed to evaluate the presence of significant differences among the four configurations. In addition, results are qualitatively evaluated by looking at class activation maps (CAMs) produced by each model, to assess the quality of the features maps.

Figure 1 shows the results of E1, E2, E3 and E4 when processing Z-brain. E2 showed the lowest performance

for all brain planes and with all the considered percentages of training patients. We conjectured this behaviour is due to the challenges of performing self-supervision from a dataset (i) with low inter-class variability and (ii) small in size (1527 images in total). E1 showed good performance, reinforcing the concept that knowledge obtained from a bigger dataset could also be beneficial in a medical image contest [4]. E3 and E4 showed consistently high F1-score for all brain planes and all the percentage of training images. Good performances are always reached when considering TC, having the plane unique characteristics (e.g. presence of cerebellum and cisterna magna) as opposed to TT and TV. As regards TT and TV, since these planes are very close to each other, it is more likely for them to be misclassified one with the other. No significant differences were found when comparing E1, E2, E3, and E4 (ANOVA test p -value > 0.05) on the various percentages of training patients. However, this is not in accordance with the qualitative results obtained with CAM. It is in fact clear from Figure 2 how E4 experiment influences the way the architecture looks at the image in order to make a prediction.

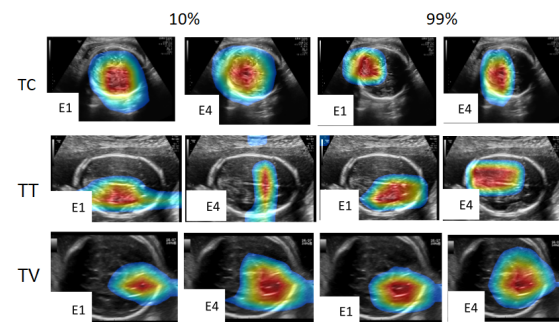


Figure 2: Visual samples of CAMs obtained from E1 and E4 experiments considering 10% and 90% of annotated patients. Compared to E1, E4 seems to focus on the most discriminative landmarks during the decision making process.

¹<https://zenodo.org/record/3904280>

3. Stenosis detection from coronary angiography

Coronary angiography is a medical imaging technique, that uses X-rays and contrast dye to visualize the coronary arteries: in practice, a thin catheter is inserted into an artery and it is used to inject a radiocontrast agent, that reveals arteries structure on X-ray images [5]. This procedure is commonly used to diagnose and guide treatment for coronary artery disease (CAD), which is caused by the presence of stenosis, narrowing of coronary arteries due to plaque buildups possibly leading to heart attack and other cardiovascular complications. Stenosis can be detected through this procedure by identifying areas where the contrast dye is impeded or delayed in its flow, however the accuracy of this procedure depends on the expertise of the clinician, which needs to own a detailed knowledge of normal coronary arterial anatomy and its common variants [6]. To provide a more objective and reliable assessment of coronary artery stenosis reducing inter-observer variability, computer-based methods are suited. More traditional computer-aided systems use machine learning algorithms to automatically identify stenotic lesions in the coronary arteries involving several steps: preprocessing (i.e. enhancement of the angiographic images to improve vessel visibility), vessel segmentation, relevant features extraction (such as vessel diameter, tortuosity, and stenosis severity), and classification (vessel classified as stenotic or not based on the extracted features). Recent studies [7, 8, 9, 10] have shown that DL algorithms can achieve promising results in stenosis detection from coronary angiography, even though they have to deal with many challenges: for instance, the poor contrast among vessels and background, the variability of the coronary artery anatomy and the appearance of stenotic lesions. To address these challenges, DL algorithms should be trained on large datasets of angiographic images with varying levels of stenosis severity, with images from different patients with varying anatomical variations and annotated from several experts.

Hence, to tackle this problem, we are working on a federated DL approach using data from two different medical centres to improve the accuracy and generalizability of stenosis detection models while preserving data privacy. We collected 1566 coronary angiography images from 245 patients (Dataset A) at the Ospedali Riuniti of Ancona (Italy), annotated by three different clinicians; at the same time a dataset of 8325 coronary angiography images of 100 patients (Dataset B) reviewed by one operator has been made available from Danilov et al. [8]. As primary step, before adopting a federated unified approach we trained and tested state-of-the-art architectures with each dataset and we thoroughly ex-

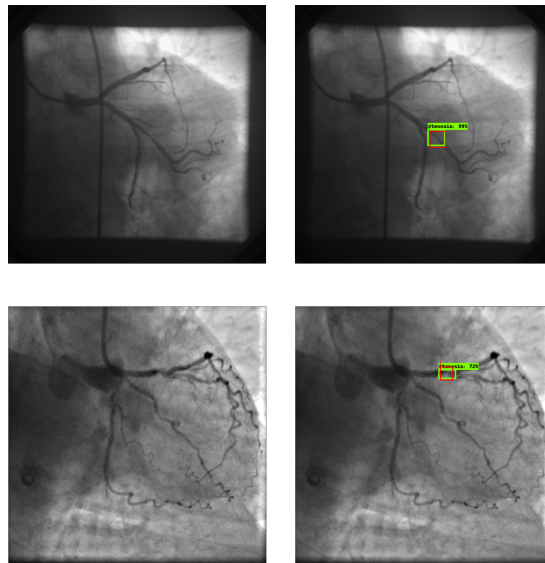


Figure 3: Samples of coronary angiographic images from Dataset A (first row) and Dataset B (second row) with the associated prediction (ground truth bounding box in red and predicted bounding box with confidence score in green).

plored distribution and similarity of the two datasets. The performance achieved by training state-of-the-art models with each dataset are evaluated and the best results are achieved with a Faster RCNN architecture reaching a F1-score value of 0.63 for testing set of Dataset A (112 images of 25 patients) and 0.84 for testing set of Dataset B (830 images of 3 patients). An example of stenosis detection is reported in Figure 3. These preliminary outcomes result to be in line with current literature [8, 9, 10], however they show strong differences strictly dependent on data characteristics. In fact, the two datasets have noticeable differences, and Dataset A is particularly complex because it includes fewer images from more patients, which results in higher variability; additionally, it has been annotated by multiple operators and the images are noisier due to the presence of distracting objects like suture points. As a result, the use of domain adaptation techniques is being considered.

Overall, DL techniques have already shown great potential in this field, and future developments are expected to further improve the accuracy and provide an efficient and general support for CAD diagnosis, ultimately leading to better patient outcomes.

4. Cancer detection from narrow-band laryngoscopy

While promising results have been obtained in several image modalities, the analysis of videoendoscopic frames still represents a challenge [11]. This may be explained considering the peculiar challenges of endoscopic videos, including poor contrast, low signal-to-noise ratio, presence of motion blurring, and tissue motion. DL approaches applied to video-analysis are of particular interest in the field of head and neck oncology, given that endoscopic examination is a crucial step in diagnosis, staging, and follow-up of patients affected by upper aerodigestive tract (UADT) cancers. Instance segmentation is particularly suited to the context of UADT endoscopy since different alterations (e.g., concomitant inflammatory or benign lesions) can be frequently encountered in the field of view together with the target lesion, and due to the fact that patients with head and neck cancers can develop distinct islands of neoplastic or dysplastic mucosa (i.e., field of cancerization) that might involve various portions of the videoframe, even without continuity.

Our approach to UADT cancer segmentation makes use of a Mask R-CNN [12], which consists of a backbone, a Region Proposal Network (RPN), and three heads for classification, bounding-box regression, and segmentation. As backbone, we used a ResNet50 pretrained on the COCO dataset combined with the Feature Pyramid Network (FPN), to extract features from the input frame at multiple scales. Starting from the features computed with the backbone, the RPN identifies candidate regions containing the tumor. For each of the proposed regions, the final bounding box containing the tumor and the tumor segmentation are obtained from the three heads.

The study was performed including videoendoscopies acquired from 323 patients treated at the Unit of Otorhinolaryngology - Head and Neck Surgery, University of Brescia, Italy for UADT cancer. A total of 1034 videoendoscopic frames was selected from a dedicated archive and anonymized. Three different subsets were generated according to the lesion primary site: oral cavity, oropharynx, and larynx/hypopharynx. To train and test the algorithm, the dataset was further split over patients balancing the three classes into three sets: 935 images from 290 subjects for training, 48 images from 16 subjects for validation, and 51 images from 17 subjects for testing.

The tumor segmentation performance was measured using the Dice similarity coefficient (*DSC*) and other spatial overlap-based metrics as accuracy (*Acc*), recall (*Rec*), specificity (*Spec*), precision (*Prec*) and intersection over union (*IoU*) which achieved the following values when computed on the test set: $DSC = 0.79 \pm 0.23$, $Acc = 0.91 \pm 0.12$, $Rec = 0.91 \pm 0.22$, $Spec = 0.93 \pm 0.12$, $Prec = 0.85 \pm 0.24$, $IoU = 0.73 \pm 0.27$.

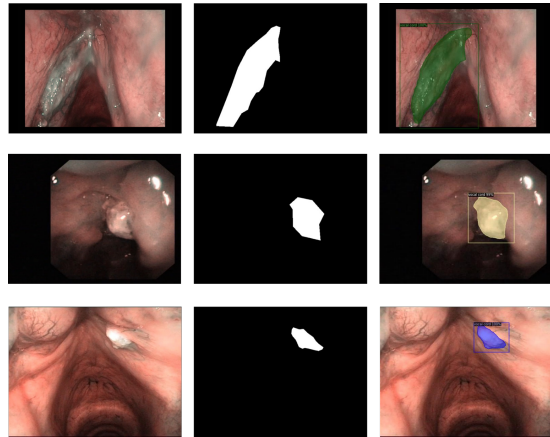


Figure 4: Visual samples of the segmentation results. From left to right: raw endoscopic frames, ground truth annotations, and predictions obtained with the proposed method.

Sample segmentation results are shown in Fig. 4 to visually compare the results of the proposed model with the ground truth. This study includes three sites of the UADT to allow a comparison of the algorithm’s diagnostic performance in different anatomical areas. The algorithm was able to identify and segment the lesion in 76.5% of cases, and showed remarkable diagnostic accuracy. Interestingly, results were significantly inferior in the oral cavity, where all outcome measures underperformed when compared with larynx/hypopharynx and, in some cases (in terms of accuracy), oropharynx. This result is possibly related to the wide variety of epithelial subtypes observed in the oral cavity, which produce additional complications to the oral examination (e.g., presence of light artifacts), and confounding factors (e.g., tongue blade, teeth, or dentures) that the DL algorithm must learn to take into account.

5. Infants’ monitoring with depth images

The assessment of the quality of preterm infants’ spontaneous motility is recognized as a highly reliable tool to early diagnose future neuro-motor impairments, which premature children develop much more frequently than the rest of the population [13, 14, 15]. Despite its recognized clinical relevance and reliability, the diffusion of this assessment is currently hindered by its high economical and temporal requirements. In fact, the clinical evaluation of the quality of these movements is entrusted to highly-trained clinicians who need to monitor the infant for a sufficiently long time. Consequently, this practice emerges as time-consuming, discontinuous, and

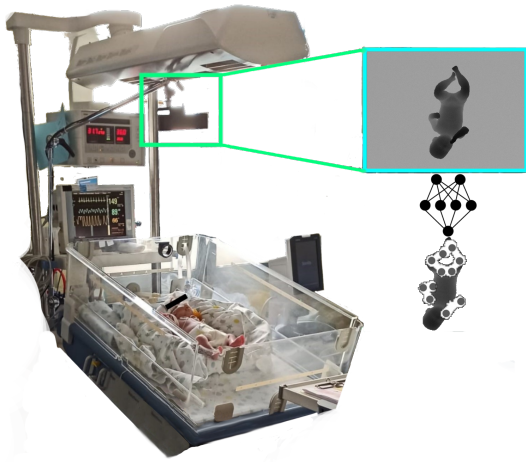


Figure 5: Acquisition set-up in a Neonatal Intensive Care Unit, and example of limb-pose estimation from a depth-video frame via BabyPoseNet, the segmentation CNN described in [17].

expertise-dependent, and the final diagnosis is extremely subjective and prone to fatigue bias, as well as to inter-subject variability.

To overcome these issues, to support clinicians in this delicate assessment, and to spur the diffusion of this important practice, our research group works on the design and development of a DL-based system to automatically monitor preterm infants in neonatal intensive care units (NICUs). In particular, the system relies on an (RGB-)D (RGB and depth) camera to be placed on top of a crib and to constantly monitor the infant. The choice of using depth images was driven by the necessity to protect the privacy of the people involved, whether it's the children or their parents, or the clinical personnel who interacts with the infant. This allowed us to collect a dataset (the BabyPose dataset [16]) that we could make publicly available without any privacy-related concern.

Our first approach [17] involved two cascaded CNNs, a U-Net-based segmentation network and a regression network to refine the predictions of the positions of the anatomical regions of interest (the 12 joints of the limbs and the 8 connections between them). In particular, the segmentation network (BabyPoseNet) features a bi-branch architecture, which means that every stage of the CNN process the information in two parallel branches, and concatenates the two thus-obtained processed data back into one single tensor. This choice is driven by the fact that the two families of entities that the network

needs to localize (joints and joint-connections) are very different, yet belong to one body. As can be seen in Fig. 5, this first method showed promising results (average $DSC = 0.80$, average $Rec = 0.70$), so the second step was to improve the prediction performance by using the temporal continuity of the movements. Hence, the second approach [18] involved the use of 4-dimensional tensors, *i.e.*, video batches of consecutive frames. The new temporal information improved the performance: DSC values improved by 0.03/1, and the Rec values improved by 0.07/1. However, due to the bi-branch structure and the 4-dimensional data handling, this second framework is particularly expensive in terms of computational requirements, which usually entail more expensive hardware. Since these technologies are intended for healthcare, a sector which is notably subject to inequalities, developing expensive frameworks is an unfair practice.

Therefore, following the principles of Green AI [19], we chose to develop fairer technologies, that could perform well enough to be used and relied on by the healthcare sector, but that require less computation (and that are, henceforth, less expensive) than the existing approaches in state of the art. This line of research led us to modify the previously designed 2D segmentation CNN (BabyPoseNet), to obtain a new architecture (TwinEDA [20]) that exploits lightweight approximations of otherwise computationally-intensive traditional convolutions (like asymmetric and dilated convolutions) to reduce the computational load required to process the data. TwinEDA is twice as fast as BabyPoseNet when processing single frames, requires less than half the number of Floating Point Operations, but its performance is totally comparable with that of BabyPoseNet, both in terms of DSC and Rec . Current research involves the use of Knowledge Distillation, a technique to transfer the knowledge learnt by a big CNN into a smaller and/or more shallow one, without significant accuracy losses.

Additionally, we are also carrying out research on the ethical aspects related to the use of DL in the actual clinical practice, especially when the practice involves infants and their images. Thanks to the collaboration with artificial intelligence ethicists, and clinicians from the Salesi Hospital (Ancona, Italy), we identified the most probable ethical issues that might arise from the application of this kind of technologies to a clinical environment. We outlined a framework to help DL researchers design technologies that are trustworthy and accountable by design, which is of crucial importance to overcome the very widespread mistrust towards artificial intelligence, especially when applied to the healthcare domain.

6. Conclusion

This contribution summarized our most recent research work in the field of medical image analysis. We plan to push the state of the art in the field forward by working on (i) semi, weak and self-supervised learning to attenuate the issue of having small annotated datasets, (ii) model efficiency, to reduce the energy consumption and CO2 emission when training and deploying our algorithms and be aligned with the European Green Deal, (iii) federated learning, for data-private multi-institutional collaborations, where model-learning leverages all available data without sharing data between institutions and (iv) adherence to the ethics guidelines for trustworthy AI and to the AI act for what concerns legal aspects of AI.

References

- [1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, *Medical Image Analysis* 42 (2017) 60–88.
- [2] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, S. Moccia, A review on deep-learning algorithms for fetal ultrasound-image analysis, *Medical Image Analysis* (2022) 102629.
- [3] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 1597–1607.
- [4] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning?, *IEEE transactions on medical imaging* 35 (2016) 1299–1312.
- [5] V. Gorennoi, M. P. Schonermarck, A. Hagen, Ct coronary angiography vs. invasive coronary angiography in chd, *GMS health technology assessment* 8 (2012).
- [6] W. Grossman, *Grossman’s cardiac catheterization, angiography, and intervention*, Lippincott Williams & Wilkins, 2006.
- [7] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, J. Ruiz-Pinales, Improving convolutional neural network learning based on a hierarchical bezier generative model for stenosis detection in x-ray images, *Computer Methods and Programs in Biomedicine* 219 (2022) 106767.
- [8] V. V. Danilov, K. Y. Klyshnikov, O. M. Gerget, A. G. Kutikhin, V. I. Ganyukov, A. F. Frangi, E. A. Ovcharenko, Real-time coronary artery stenosis detection based on modern neural networks, *Scientific reports* 11 (2021) 1–13.
- [9] K. Pang, D. Ai, H. Fang, J. Fan, H. Song, J. Yang, Stenosis-detnet: Sequence consistency-based stenosis detection for x-ray coronary angiography, *Computerized Medical Imaging and Graphics* 89 (2021) 101900.
- [10] W. Wu, J. Zhang, H. Xie, Y. Zhao, S. Zhang, L. Gu, Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint, *Computers in biology and medicine* 118 (2020) 103657.
- [11] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni, et al., Surgical data science—from concepts toward clinical translation, *Medical Image Analysis* 76 (2022) 102306.
- [12] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
- [13] T. Wardlaw, D. You, L. Hug, A. Amouzou, H. Newby, Unicef report: enormous progress in child survival but greater focus on newborns urgently needed, *Reproductive health* 11 (2014) 1–4.
- [14] C. Einspieler, H. F. Prechtel, F. Ferrari, G. Cioni, A. F. Bos, The qualitative assessment of general movements in preterm, term and young infants—review of the methodology, *Early Human Development* 50 (1997) 47–60.
- [15] M. Porro, C. Fontana, M. L. Gianni, N. Pesenti, T. Boggini, A. De Carli, G. De Bon, G. Lucco, F. Mosca, M. Fumagalli, et al., Early detection of general movements trajectories in very low birth weight infants, *Scientific Reports* 10 (2020) 1–7.
- [16] L. Migliorelli, S. Moccia, R. Pietrini, V. P. Carnielli, E. Frontoni, The babypose dataset, *Data in Brief* 33 (2020) 106329.
- [17] S. Moccia, L. Migliorelli, R. Pietrini, E. Frontoni, Preterm infants’ limb-pose estimation from depth images using convolutional neural networks, in: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2019, pp. 1–7.
- [18] S. Moccia, L. Migliorelli, V. Carnielli, E. Frontoni, Preterm infants’ pose estimation with spatio-temporal features, *IEEE Transactions on Biomedical Engineering* 67 (2020) 2370–2380.
- [19] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green AI, *Communications of the ACM* 63 (2020) 54–63.
- [20] L. Migliorelli, A. Cacciatore, V. Ottaviani, D. Berardini, R. L. Dellaca, E. Frontoni, S. Moccia, Twineda: a sustainable deep-learning approach for limb-position estimation in preterm infants’ depth images, *Medical & Biological Engineering & Computing* 61 (2023) 387–397.