


Deep-learning-based markerless tracking of distal anatomical landmarks in clinically recorded videos for assessing infant movement patterns associated with neurodevelopmental status*

Hamid Abbasi ^a, Sarah R. Mollet^a, Sian A. Williams^{b,c}, Malcolm R. Battin^d, Thor F. Besier^a and Angus J. C. McMorland^{a,e}

^aAuckland Bioengineering Institute (ABI), University of Auckland, Auckland, New Zealand, New Zealand; ^bLiggins Institute, University of Auckland, Auckland, New Zealand; ^cCurtin School of Allied Health, Curtin University, Perth, Australia; ^dDepartment of Newborn Services, Auckland City Hospital, Auckland, New Zealand; ^eDepartment of Exercise Sciences, Faculty of Science, University of Auckland, Auckland, New Zealand

ABSTRACT

Abnormal patterns in infants' General Movements (GMs) are robust clinical indicators for the progression of neurodevelopmental disorders, including cerebral palsy. Availability of automated platforms for General Movements Assessments (GMA) could improve screening rate and allow identifying at-risk infants. While we have previously shown that deep-learning schemes can accurately track the longitudinal axes of infant limb movements (12 anatomical locations, 3 per limb), information about the distal limb segments' rotational movements is important for making an accurate clinical assessment, but has not previously been captured. Here we show that training schemes that are highly successful at tracking trunk and proximal limb landmarks perform less well for the distal limb landmarks, and this problem is exacerbated when landmarks are more precisely defined in the training-set to capture rotational movements. Increasing the sample size to 26 videos using a mixture of laboratory and clinical data pre-selected for diversity of pose and video conditions in a ResNet-152 deep-net model was sufficient to permit accuracy of >85% for the distal markers, and overall accuracy of 98.28% (SD 2.29) across the 24 landmarks. This scheme is suitable to form the basis of an infant pose reconstruction algorithm that captures clinically relevant information for an automated GMA.

ARTICLE HISTORY

Received 15 March 2023
Accepted 4 October 2023


KEYWORDS

Automated markerless motion tracking; clinical recordings; deep learning; DeepLabCut; image processing; motion capture; neonatal General Movements Assessment (GMA)

Introduction

Neurological injuries due to hypoxic-ischemic encephalopathy, perinatal stroke and infection are life-threatening conditions to a developing fetal or neonatal brain. Such injuries can often lead to neurodevelopmental impairments that cause lifelong physical

CONTACT H. Abbasi  h.abbasi@auckland.ac.nz

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/03036758.2023.2269095>.

* This paper was an invited article from Royal Society Te Aparangi in recognition of Hamid Abbasi receiving the Cooper Award at the 2022 Research Honours Aotearoa.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

disability, including cerebral palsy (CP) (Fairhurst 2012; Gunn and Thoresen 2015; Ahearne et al. 2016; Abbasi et al. 2023). Spontaneous General Movements (GMs) of infants during the first 6–20 weeks of post-term age contain valuable prognostic information about the quality of neurodevelopmental growth (Ferrari et al. 2004; Garcia et al. 2004; Spittle et al. 2009; Bosanquet et al. 2013). The complexity and variations of GMs are thought to be driven by cortical structures modulating brainstem and spinal cord central pattern generator networks (Hadders-Algra 2018; Prechtl 1997). Consequently, abnormal or absent GMs can indicate risks of impaired neurodevelopment or neuromotor deficits (Morgan et al. 2019; Prechtl 1997). Clinically, the General Movements Assessment (GMA) is used to identify infants with higher risk of neurodevelopmental disorders, such as cerebral palsy (CP). Cramped-synchronised GMs with rigid and stiff features during what is termed the ‘writhing period’ (up to around 8 weeks post-term age) as well as the absence of fidgety movements in the ‘fidgety period’ (12–20 weeks post-term age) are prognostic signatures of possible developing CP (Einspieler and Prechtl 2005; Darsaklis et al. 2011). Early diagnosis provides an opportunity for early access to interventions while the brain’s neuroplasticity is still high, which can improve neuromuscular outcomes (Hadders-Algra 2014; Bernava et al. 2022; Caruso et al. 2020). The widespread utility of the GMA is currently limited at paediatric clinics, in part due to the limited number of qualified assessors; automated computer-vision and classification technology can alleviate this limitation.

Advanced image processing techniques, including deep-learning technology, have recently created new robust motion-tracking algorithms to capture body motions from videos. Toolboxes such as DeepLabCut™ (Nath et al. 2019; Mathis et al. 2018) and OpenPose (Cao et al. 2019), are good examples of motion tracking platforms that take advantage of these principles. DeepLabCut (Mathis et al. 2018) in particular has shown robust capabilities for markerless pose-estimation and movement tracking of various species including non-primates (Mathis et al. 2021) and primates (Labuguen et al. 2021; Abbasi et al. 2023). Deep-net models developed in DeepLabCut are shown to be capable of identifying previously learned landmarks and track movements in out-of-domain human (Abbasi et al. 2023) and animal subjects (Mathis et al. 2021), as well as human infants (Nath et al. 2019; Wei and Kording 2018). DeepLabCut has also been shown to effectively generalise when tested on out-of-training domain subjects.

DeepLabCut is built on transfer-learning through using ResNet models pretrained on a benchmark ImageNet (Krizhevsky et al. 2012) for object detection, and utilises deconvolutional layers for semantic segmentation (Mathis et al. 2018). Deconvolutional layers are generally designed for up-sampling the visual information to estimate probability of an object similarity to a previously learned object, in spatial space, and localise a body-part in an image. These features seem to have helped DeepLabCut achieve reliable motion tracking performance. Deeper network structure in deep-net models (e.g. ResNet-152) has been helpful for mimicking the mechanism of biological neuronal cells and their complex connectivity (Krizhevsky et al. 2017). ResNet-152, in particular, is one of the enhanced deep-net models which was first introduced by Microsoft in the 2016 ImageNet challenge (Image-net Challenge n.d.). During the training process, the manually labelled data are iteratively used for fine-tuning the deep-net model’s weight parameters towards minimising the value of cross-entropy loss function. The scheme initially assigns high probabilities to human-labelled

locations in the manually annotated images while the rest of the image gains negligible likelihoods (Mathis et al. 2018).

The application of deep-learning-based approaches is relatively new in the infant GMA field. Some recent attempts have applied these techniques to movement classification, not motion detection in 2D videos (Cunningham et al. 2019; Lempereur et al. 2020; Silva et al. 2021). Other recent works have demonstrated capabilities of deep-learning-based approaches for prediction of CP and other neurological disorders in high-risk infants from automated markerless tracking of their GM patterns (Groos 2022a; Sakkos et al. 2021; Raghuram et al. 2022; Wu et al. 2022; Shin et al. 2022; Groos et al. 2022b). Several teams (Groos et al. 2022a, 2022b; Wu et al. 2022; Shin et al. 2022), including us (Abbasi et al. 2022)[27], have used deep-learning-based markerless motion tracking technology to track supine infant movements equivalent to the movements recorded and assessed during the GMA. All of these studies have used deep learning to track and reconstruct three markers on each arm and leg, giving the longitudinal axis of the limb. In our case, we showed that, using deep-net models trained in the DeepLabCut™ environment (Nath et al. 2019; Mathis et al. 2018), we could achieve a cross-validated accuracy of 95.52% (SD 2.43) across 12 anatomical locations when trained on data from only five infants and tested on out-of-domain videos from a novel baby.

The original descriptions of GMs made by Prechtl and others specifically mention the importance of limb rotations: ‘The majority of extension or flexion of arms and legs is complex, with superimposed rotations and often slight changes in direction of the movement. These additional components make the movement fluent and elegant and create the impression of complexity and variability’ (Prechtl 1990). Similarly, GMA-trained clinicians state that small rotations are a key feature of the fidgety movements being assessed. However, to date no studies have used deep learning to track more than three markers per arm, as would be required to capture these rotational movements. It is possible that missing this feature of infant movements restricts automated analyses from identifying important signs relating to neuromotor development.

This current work assesses the implications of tracking additional, more precise anatomical landmarks allowing the estimation of distal limb segment rotations that have been beyond the scope of previous works, and how tracking performance of those distal landmarks is affected by using larger training datasets. We assess the generalisation capabilities of two landmark sets for training of the deep-net, one tracking 16 anatomical landmarks and the other 24 landmarks including three differently positioned points on each hand and foot compared to the 16-landmark set. Generalisation is assessed for models trained with both (1) a clinically recorded dataset, and (2) a combined dataset of clinical and laboratory videos. Leave-one-out cross-validation (LOOCV) was used to assess the trained models’ performance, across all landmarks, across the entirety of three randomly selected videos from the clinical cohort, using our previously reported Kalman filter-based assessment approach.

Data acquisition

Ethics

All procedures in this study were approved by the Auckland Health Research Ethics Committee (AHREC-000146). Parents/caregivers were fully informed of the purpose

of experiments, filming procedure and methods, and provided informed written consent for their child's participation.

Recording procedures

Clinical recordings were provided by the Child Development Centre of the Waikato District Health Board (DHB), New Zealand. Infants were filmed by a trained clinician during March 2017 to June 2021 as part of clinical practice. Data from a cohort of 15 infants (9 males and 6 females, born with gestational ages ranging 24–41 weeks, mean age when recorded 22.9 (SD 4.1) weeks, mean corrected age when recorded 12.8 (SD 1.0) weeks; Table 1) with GMs captured during their fidgety period were selected from the larger database of available clinical recordings to provide a diverse selection of movement patterns, skin colour, clothing and background colour and videoing conditions such as shadowing and angle. Selection criteria for the larger database were set to infants with less than 32 weeks gestation and/or less than 1500 g weight at birth. Also, an HIE (Sarnat score 2 or 3) or any other concerns identified by medical staff was considered for infant selection. Infants were filmed either at home or brought to the clinic, changed into a nappy only, and laid down in a supine position on a plain-colour linen on a comfortable mat. Standard iPads (MQDT2X/A: 12-megapixel camera with 4K HD video and a MD367X/A: 3rd generation iPad with 5-megapixel 1080p HD camera) were used for recording the videos. Infants were filmed while awake, with spontaneous mobility, in their natural state for 1–2 minutes with minimal environmental distractions (i.e. from parents/caregivers or the presence of toys). Data were recorded with 1920 × 1080 pixels resolution and frequency of 29.97 frames/second. Data were initially post-processed in Adobe Premiere Pro 2022 to remove intervals where parts of the infant's body were out of frame, placing the infant relatively in the center of the recording and further cropping out any excessive background space in the video. Based on experience, removing the extra unnecessary information in the videos could be helpful to improve the robustness of the model. Videos were then saved in their original quality. A total of 24K frames were eventually used from the 15 infants.

Table 1. Infants' health and demographic information. All participants belong to a high-risk group, had birth complications and comorbidities and received medications and/or continuous positive airway pressure (CPAP).

	<i>Gender</i>	<i>Ethnicity</i>	<i>Gestational age at birth (wks + days)</i>	<i>Ventilated</i>
P01	M	Māori	26 + 0	Yes
P02	M	NZE [^]	27 + 6	Yes
P03	M	Middle Eastern	32 + 0	No
P04	F	NZE	29 + 6	No
P05	M	Māori	24 + 0	Yes
P06	F	Māori	27 + 0	No
P07	F	Māori/NZE	30 + 2	No
P08	M	Māori Pacific	28 + 0	No
P09	M	African	41 + 0	No
P10	F	NZE	35 + 0	No
P11	M	NZE	30 + 3	Yes
P12	F	Māori	26 + 5	Yes
P13	F	NZE	33 + 0	No
P14	M	NZE	26 + 0	Yes
P15	M	NZE	30 + 5	No

[^] New Zealand European

The laboratory dataset comprised 12 videos from 6 term-born infants with mean age when recorded of 17.33 (SD 2.9) weeks, 4 male:2 female. Further details of the laboratory-based recordings are provided in Abbasi et al. 2023.

Methods and computational approach

Deep-learning training and validation

A generalised motion tracking deep-net model is important for our application, where we would need the network to be able to reliably identify and track body landmarks in novel videos from unseen infants. To evaluate how the tracking performance for the distal limb segments varies between the two landmark sets, we randomly selected 3 infant videos from the pool of 15 clinically recorded infants. To assess the sensitivity of tracking performance to the training set size, four training sets were then formed, schemes #1–#3 included 100 frames from each of 5, 10 or 14 videos from the clinical dataset, respectively. Scheme #4 was formed by adding 100 frames of each of 12 laboratory-recorded videos to the training data of scheme #3. The 100 frames from each video were selected using DeepLabCut's k-means clustering feature to maximise diversity of postures in the selected frames.

A total of 64,800 points (clinical recordings: 15 infants \times 100 frames each \times 24 anatomical locations = 36,000; and lab recordings: 6 infants \times 2 videos \times 100 frames each \times 24 anatomical locations = 28,800) were manually labelled by an expert (H. A.). Examples of the manually labelled locations in the clinically recorded videos are shown with coloured dots in Figure 1. This figure also illustrates the diversity of ethnicities, filming conditions and quality of the videos used from the clinical recording cohort.

We then assessed, using cross-validation, whether models trained with each of the four schemes performed equally well across the three left-out subjects. For testing schemes #1 and #2, none of the training sets included any of the three testing videos. For testing schemes #3 and #4, one video from the clinical dataset was selected for testing and the training set was formed from the remaining 14 clinical videos (plus the laboratory videos for scheme #4).

We used DeepLabCut's recommended multi-step learning-rate updating regime (0.005 from 0–10k iterations, 0.02 for 10k–430k iterations and 0.002 for 430k–700k iterations) and their built-in 'imgaug' image augmentation algorithm with a training-to-test proportion of 95% to 5% for each training round. A decaying cross-entropy loss function across all training schemes confirmed validity of the training process. Depending on the size of the training dataset in each scheme, network training, which was run for 700k iterations, took 3 to 7 days on the NVIDIA A100 machines (detailed in the 'Computing Infrastructure' section). The root mean square errors (RMSE), generally the average distance between the detected labels by the ResNet and the scorer's annotations, were 2.44 (SD 0.10) pixels after training with the 24-landmark set, while test errors were found to be higher at 4.17 (SD 0.05) pixels. Since physical calibration scales were not included in the clinical data set, and imaging conditions varied between videos, we are unable to convert these pixel values to physical distances. Plots of cross-entropy loss for all four training schemes after training are shown in Figure 2, confirming the fast convergence



Figure 1. Examples of the 24 manually labelled anatomical locations (colourful dots), in DeepLabCut environment, in 15 clinically recorded infants. No consent was received to share baby #13’s image.

of the model across all training schemes. The discontinuity at 430,000 iterations is a result of the standard-practice parameter change in the learning-rate scheme from the 0.02 to 0.002.

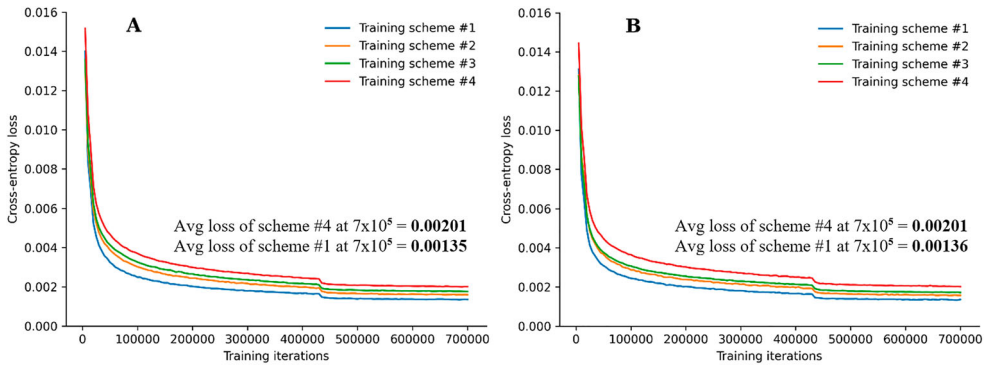


Figure 2. Cross-entropy loss for all training schemes from **A**, 24 anatomical landmarks and **B**, 16 anatomical landmarks.

Labelling strategies

The 16-landmark set comprised labels at the eyes, nose, sternum and three locations per limb: shoulders, elbows and hands, or anterior superior iliac spines (ASIS), knees and feet. All landmarks were labelled at the 2-D position of their location in the image, regardless of the orientation of the structure (Figure 3 A–D). For example, the hand label was placed on the dorsal, palmar, ulnar or radial aspect of the hand depending on the orientation of the hand, such that the label was above the point deep in the hand below the center of the palm. The only time a landmark was not labelled in an image was when the landmark was covered beneath a different body part.

The 24-landmark set used the same facial, torso, shoulder and elbow landmarks as the 16-landmark set. The exact same manual label locations were used from the 16-landmark data. Hands and feet were labelled differently, with a label placed at the wrist, index finger metacarpophalangeal (MCP) joint, little finger MCP, ankle, big toe metatarsophalangeal joint (MTP) and little toe MTP. As before, landmarks were labelled regardless of orientation, at the apparent location of the central point deep within the anatomical structure, except when covered by a different body part, where they were not labelled. Landmarks on the MCPs and MTPs were not labelled when that side of the hand was not visible in the image because of the hand orientation (Figure 3 E–H).

Automatic unsupervised performance measure

We have previously described the development of a Kalman-filter-based approach to automatically generate performance metrics using all frames of a validation video

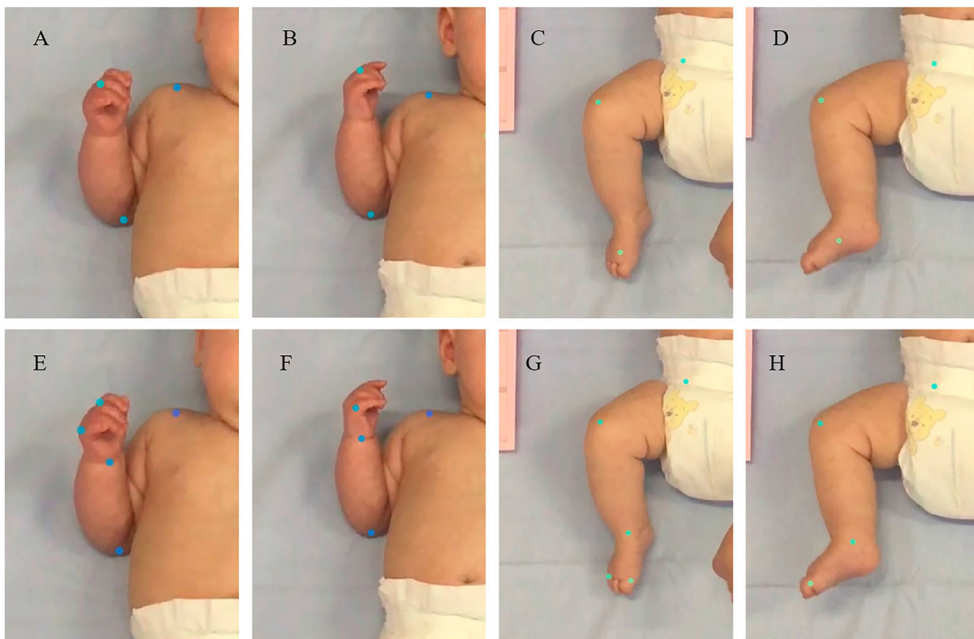


Figure 3. Examples of the manually labelled anatomical locations (colourful dots) in the right limbs from the 16-landmark set (A–D), as well as the 24-landmark set (E–H).

(Abbasi et al. 2023). This approach supplants earlier methods using the average Percentage of Correct Keypoints (aPCK) (Liu et al. 2021; Mathis et al. 2021), which is time-consuming and usually involves ‘out-of-domain’ performance validation against a much smaller number of manually labelled frames from the validation video. Briefly, the Kalman filter tracks each landmark with a state variable considering x and y position, velocity and acceleration, and this Kalman filter is run forwards and backwards through the data. The Kalman filter output is used solely to identify instances in the DeepLabCut output where the marker trajectories deviate from those smooth assumptions and therefore are likely to be labelled incorrectly. This approach helps to detect potential ‘outlier identifications’ by the deep-net landmark identifier, and at other points the landmark locations identified by the deep-learning model are unmodified by the Kalman filter process. Each frame’s DeepLabCut position estimate forms the noisy observations on which the state is updated. After each prediction step, logical combinations of likelihoods from both forward and backward Kalman filter measures are evaluated and compared against the probability of the current position of the marker that is estimated by the ResNet motion-tracking model to automatically assign true positive (TP), false positive (FP), false negative (FN) and true negative (TN) status for each marker position in each frame. The estimated probabilities from the ResNet indicate the confidence of a motion-tracking model with the identified location of that landmark in the image. Negative log-likelihood values greater than 20 were considered as a poor tracking observation, while a good labelling was defined when the ResNet model identified a marker with a confidence level of ≥ 0.6 . Similar to the previous work, here we evaluated the precision (positive predictive value, PPV), sensitivity (true positive rate, TPR), overall performance (average of precision and sensitivity) and overall accuracy (proportion of correct identifications $(TP + TN)$ out of all identifications $(TP + TN + FP + FN)$) for each anatomical landmark. For more information, refer to Abbasi et al. 2023.

Computing infrastructure

Processing was performed using New Zealand eScience Infrastructure (NeSI) high-performance computing facilities’ Cray CS400 cluster (NeSI, n.d.). The ResNet model was trained using enhanced NVIDIA Tesla A100 PCIe GPUs, with 40 GB HBM2 stacked memory bandwidth at 1555 GB/s per training task. Intel Xeon Broadwell CPUs (E5-2695v4, 2.1 GHz) were used on the cluster for handling the GPU jobs. In this work, we used DeepLabCut version 2.2.0.3, TensorFlow 2.5.0, CUDA 11.6 and Python 3.8.6.

Results

We used cross-validation to evaluate performance for tracking each trained model across all anatomical landmarks on the three left-out clinically recorded videos. The trained ResNet model was able to accurately track movements in out-of-domain (unseen) videos from three novel infants with an overall performance of 98.28% (SD 2.29), across all 24 locations, for the training scheme #4, where data from 14 clinical + 12 lab-recorded videos were used in the training set. This was improved from an overall accuracy of 92.21% (SD 9.28) for the training scheme #1 where data from only five clinically recorded videos were used in the training set (Figure 4 and Tables S1 & S2).

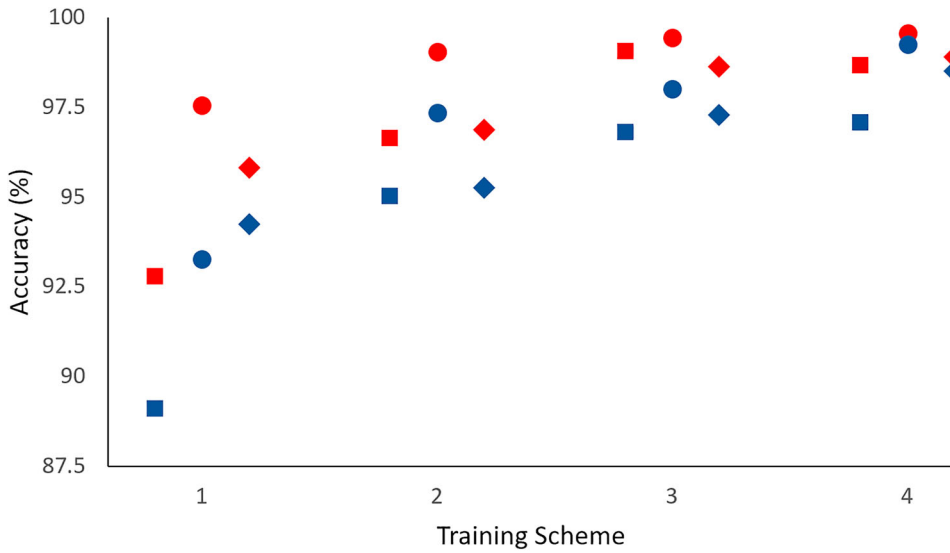


Figure 4. Accuracy values, averaged across landmarks for each testing video, for the training schemes. Red = 16-landmark set, blue = 24-landmark set; squares, circles, diamonds indicate test video 1–3 respectively.

Tabulated results in Tables S1 and S2 show how the overall performance of the ResNet model for the 16- and 24-landmark sets improves as more manually labelled data from a larger number of infants were used in the training (from training scheme #1 to training scheme #4). The performance variance also decreases as the size of the training data increases in all cases except for the change from scheme #3 to #4 in the 16-landmark set. Breakdowns of the performance by anatomical landmark are shown as heatmaps in Figure 5 (see also Tables S3–S5 and Tables S6–S10 respectively for numerical

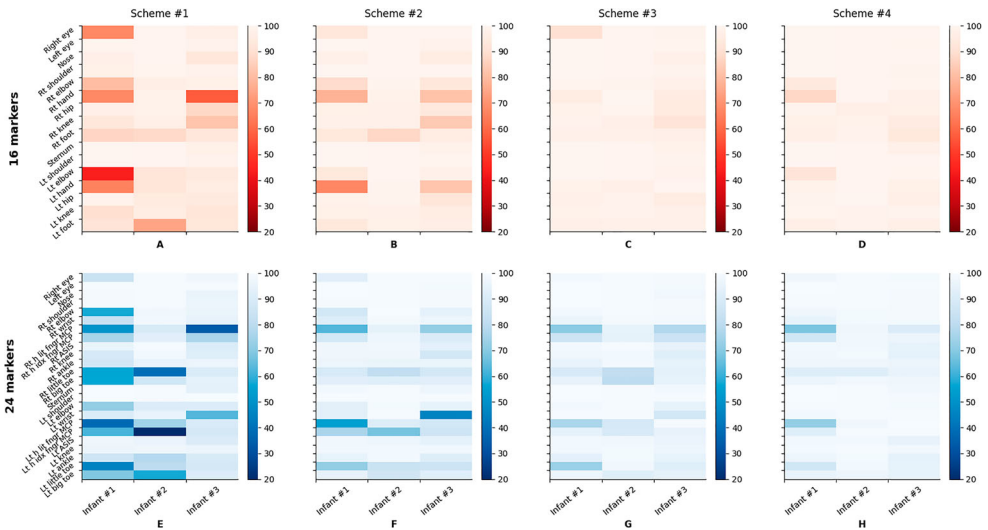


Figure 5. Accuracy measures for training scheme #1 to #4 across 16 anatomical landmarks (A–D) and 24 anatomical landmarks (E–H) on three unseen validated videos.

details). In [Figure 5 C](#) and [D](#) we can see that the increase in variance from scheme #3 to #4 with 16 landmarks is specifically caused by poorer performance in tracking the right hand of infant #1. Note that even though only the hand landmarks varied between the 16- and 24-landmark sets, model training considers all landmarks simultaneously, and therefore the models trained on the two landmark sets may have different performance even at the shared landmarks.

[Figure 6 A](#) highlights that the poorest performances occurring for the hands and elbows in the 16 landmarks with scheme #1 set recovered to much higher values for scheme #4. Similarly, the performance for MCPs and MTPs in scheme #1 of the 24-landmark set was also low and improved substantially with scheme #4 ([Figure 6 B](#)). Tables S11 and S12 provide location-based validation results across both landmark sets and all training schemes, averaged across the three test videos.

The small performance variations between test videos for training scheme #4 ([Figure 4](#)) support consistent generalisation capabilities of the deep-net model across large numbers of video-frames from three different subjects. The larger variation in the location-based comparison, especially for the 24-landmark scheme ([Figure 5 H](#), but also apparent with 16 landmarks, [Figure 5 D](#)), reflects that landmark identification is a more challenging task for the visually complex or often occluded landmarks (i.e. MCPs and MTPs).

Examples of the locations predicted by the tracking model trained with scheme #4, relative to the manually labelled locations by the expert observer (H. A.) are shown in [Figures 7 A–C](#) (16-landmark set) and [7 D–F](#) (24-landmark set).

Discussion

Previous studies in the literature developing an automated General Movements Assessment have consistently tracked three landmarks per limb. In this current work, we extend beyond that approach to a 24-landmark set which captures information about rotations of the distal limb segments. We have shown relative performance with the 16- and 24-landmark sets on a mixed dataset of clinically- and laboratory-recorded standard 2D video of infants aged 2–5 months old. The introduced set of 24 anatomical landmarks in this work, including the eyes, nose, sternum and five locations per limb is novel and clinically essential for capturing fine rotational fidgety-related movements in the hands and feet. This enhanced landmark configuration provides an advantage over the conventional three landmarks per limb, which is equivalent to the 16-landmark set scheme in this study (including eyes, nose, sternum and three locations per limb).

We evaluated the landmark identification capabilities of a deep ResNet-152 model for both of these landmark sets, which involved identification of hands and feet broadly defined in the 16-landmark set, compared to more specific limb locations such as the MCP and MTP joints within the 24-landmark set framework.

Performance impact of landmark specificity and dataset size

The performance of the tracking algorithm averaged across all landmarks consistently improved from training scheme #1 to #4 ([Figure 4](#) and Tables S1 and S2) across the three out-of-domain videos for both the 16- and 24-landmark sets. The 24-landmark

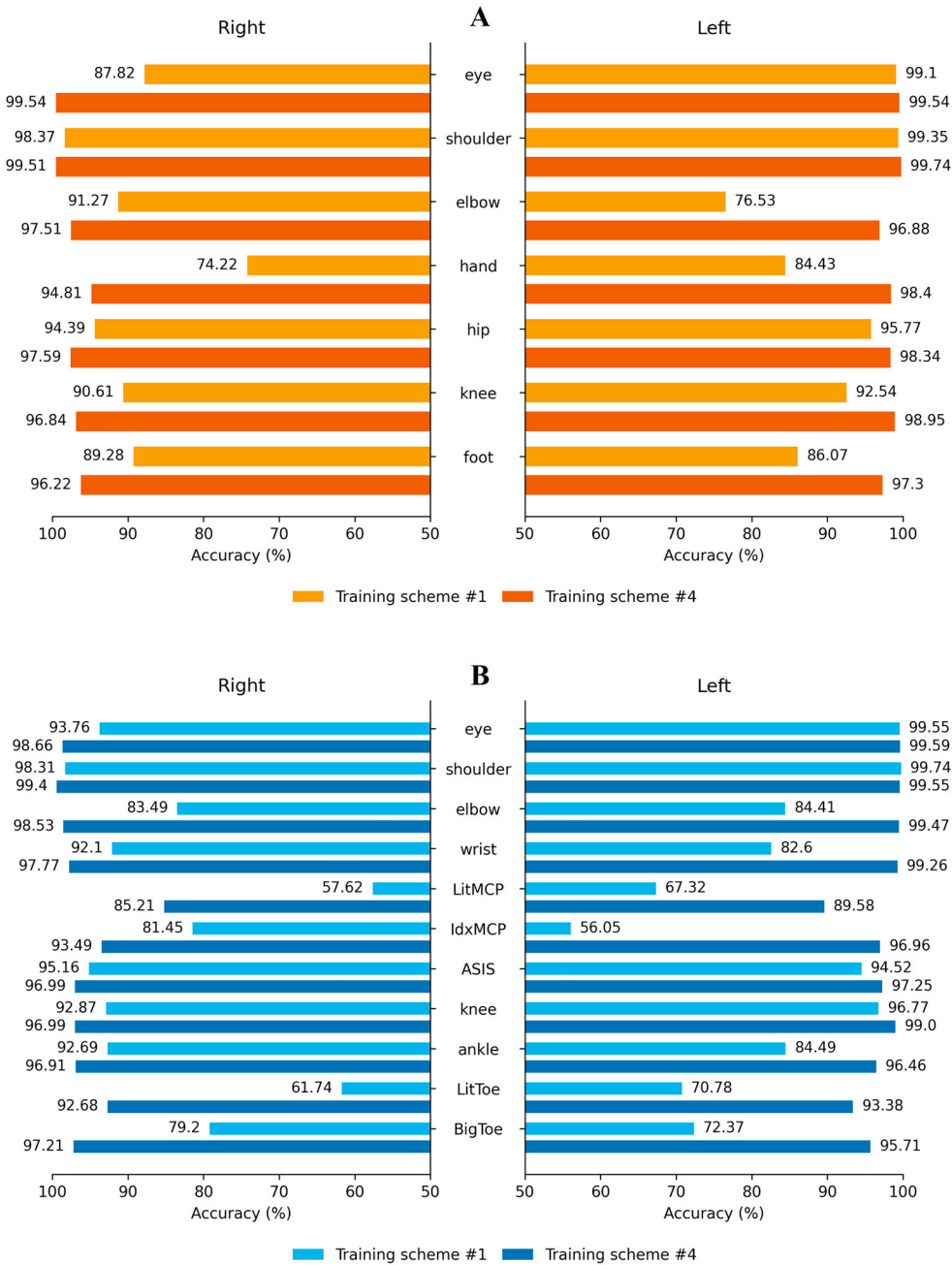


Figure 6. Location-based accuracy measures for **A**, training scheme #1 (light orange) vs #4 (dark orange) of the 16 anatomical landmarks, and **B**, training scheme #1 (light blue) vs #4 (dark blue) of the 24 anatomical landmark approach. Note: measurements for nose and sternum have been excluded (see Table S11 & S12 for details).

set, with higher distal landmark specificity, was tracked with lower performance (98.28%, SD 2.29), accompanied by increased variance, especially for the distal limbs (Figure 4 & 5) compared to the 16-landmark set (99.04%, SD 1.18).

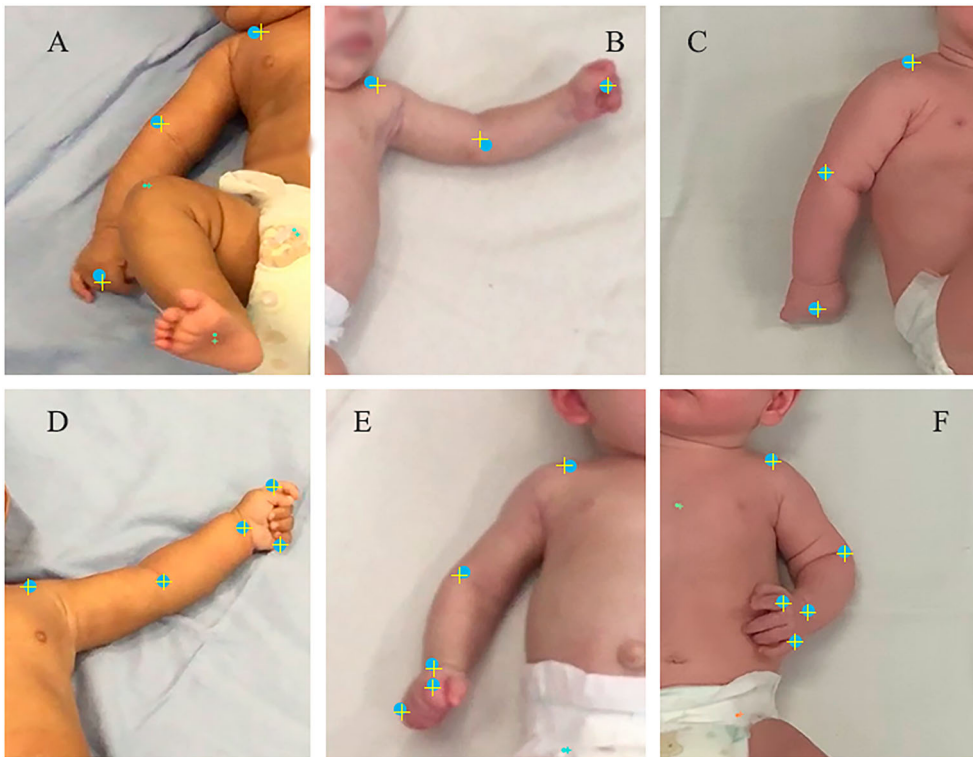


Figure 7. Examples of the automatic (+) vs manual (O) identifications in the arms of the three infants in the validation set. Results are plotted using the training scheme #4 for the 16 (A–C) and 24 landmarks (D–F). Note: marker representations of the arms are magnified for better illustration purposes.

In the 16-landmark set, the poorest performances of training scheme #1 were associated with hands, feet and occasionally one elbow, whereas tracking of these landmarks improved to similar levels to the more consistently performing landmarks when the model was trained in scheme #4 (Figure 5). The lower performance tracking of distal landmarks is exacerbated when the specificity of the landmark definition is increased (i.e. the hand MCPs and feet MTPs compared to hand and feet generally) (see Figure 4 and training scheme #1 to #4 in Figure 5). Correspondingly, the standard deviations across all anatomical landmarks in the three out-of-domain subjects were considerably and consistently lower, for all training schemes #1 to #4, in the 16-landmark set approach compared to the 24-landmark set (see Figure 4, and Tables S1 & S2).

However, our data showed that this deficiency was greatly improved when the model was trained using training scheme #4, which incorporated a larger dataset from both clinically- and laboratory-recorded videos. In fact, increasing the amount of training data to 26 videos (2600 frames), in the training scheme #4, recovered performance of the 24-landmark set back to very close to that of the 16-landmark set for two of the three test videos in training scheme #3, and above 98.5% accuracy (Figure 4).

Performance in the third test video (test video 1 in Figure 4) remains lower with scheme #4, at similar levels to that of scheme #1, suggesting that it contains novel poses not otherwise encountered in the training set.

Our evidence shows that, compared to tracking broadly defined landmarks, to achieve similar levels of performance when tracking more specifically labelled landmarks, a larger training set is required to adequately capture variations in local features within the images (Figure 4 and 5).

Unsupervised performance evaluation for large out-of-domain data

Typically, automatic performance evaluations of larger out-of-domain datasets, with thousands of frames, is a challenging task where providing ground-truth annotations through manual labelling can be very time-consuming. In this work, model performance was evaluated across all frames in the videos using our previously introduced unsupervised performance evaluation strategy that combines likelihoods of a Kalman filter and the deep-net confidence (Abbasi et al. 2023), to automatically measure performance metrics on an extensive out-of-domain recording set. Unlike the conventional aPCK approach of assessing performance on a small number of manually labelled frames (Liu et al. 2021; Mathis et al. 2021), our approach is evaluated over all frames within the video meaning that it will not miss any brief, unusual poses the infants may take, and so represents a more robust measure of performance.

Our Kalman-filter-based validation approach compares tracking results against predictions based on smooth and nearby assumptions by considering marker location, velocity and acceleration from previous frames, and combining them with the DeepLabCut confidence score to mark proper landmark identification and tracking. Related to the benefit of being able to assess tracking performance across large numbers of frames, the Kalman-filter approach is also fully unsupervised, meaning that it scales well to large datasets and allows independent identification of sections of video where the deep-net tracking has failed, presumably due to the novelty of some aspect of the pose.

The evaluated performance in this work fits well into the reported performance range of other recent markerless motion tracking studies validated over much smaller manually annotated datasets (Mathis et al. 2021).

Landmark tracking insights

The algorithm consistently demonstrated best tracking performance for eyes and shoulders, and worst for the distal landmarks (hands, MCPs and MTPs; see Figure 5).

Several factors could explain these differences. Firstly, distal landmarks can take a wider range of postures and orientations, particularly influenced by both humeral and forearm (pronation/supination) rotation, bringing different aspects of the limb to face the camera and in a range of in-plane orientations, as well as the curling or extending of fingers, which alter their appearance substantially. Features like the little finger MCP will be not visible at all, and so, unlabelled in certain configurations. Secondly, distal limbs being more mobile mean they can overlap each other, either changing their appearance while visible or occluding the hand or foot entirely from the camera. There may be fewer manually labelled frames available for these landmarks due to their occasional occlusion. However, our labelling approach does not solely focus on marking the visible landmark's location but also provides information when labelling the landmark as absent: nothing in the current image looks like this landmark. This

information should explicitly train the model to identify these cases where the landmark is not visible and implicitly help to avoid mis-identification of other body parts. Similarly, in our performance metrics, correctly identifying cases where landmarks are absent counts positively towards overall performance of the model.

An alternative explanation for the poorer performance of the distal landmarks is the increased movement speed, and therefore motion blur, of these points at a 30-frames-per-second sampling rate. We have previously identified that infant hands and feet move with similar speed distributions, while the tracking performance is poorer for hands (Abbasi et al. 2023), suggesting at least that pose variations play an important role in reducing tracking performance. Working with the standard video frame rate of 30 Hz is potentially one of the limitations of the current study. We have noticed that this sampling frequency can permit substantial motion-blur when capturing faster movements of the hands and feet. Comparison of tracking from video captured at 30 Hz and 60 Hz in a future study will allow quantification of the effect of this influence.

Enhanced landmark labelling strategy

In this study we adopted a ‘center-of-the-joint’ approach for labelling the landmarks, where the point deep in the joint was labelled whenever the joint was visible from any angle. The alternative ‘point-on-the-surface’ approach would have been to consistently label a point on the skin surface associated with a particular landmark. In this latter case, whenever that surface point was not visible, the landmark should be unlabelled. While our approach means that the visual appearance of each labelled landmark may vary more widely throughout the training set, many more labels are able to be provided, meaning that the landmark has the opportunity to be identified more often in novel data. This strategy was helpful to introduce a large diversity of infants’ poses for a specific landmark in the training set.

Dataset size and diversity

Our results show that with the three-landmark-per-arm labelling strategy used in most studies, a modestly-sized dataset of 1400 frames from 14 videos in training scheme #3 was sufficient for the ResNet-152 to achieve over 98.5% identification accuracy. For five landmarks per arm, 2600 frames from 20 infants (26 videos) was able to train the model to achieve at least 97% accuracy. Groos et al. (Groos, Adde, Støen et al. 2022) have examined the effect of training- set size between 100k and 14.5k frames for several CNN-based models (although not ResNet152). They used different Percentage of Correct Keypoints (PCK)-based performance metrics, for which equivalency with our Kalman-filter-based approach is not immediately clear. Nonetheless, their results indicate modest changes in model performance with increases in data size over 1000 frames, and that those improvements that do occur beyond this level are increases in the precision of the labelling. Two of the models reach (lower) performance maxima at 5000 frames, whereas EfficientPose-based methods continue to improve with additional data.

In the context of motion capture of infants, the effective sample size depends on the number of novel features being shown during training. If features are novel only between

infants, then the number of infants would be critical. However, in this case, the key features being trained on are poses and the appearance of body landmarks which are relatively conserved in appearance between individuals but vary substantially with different poses. It therefore makes sense that the sample size that is important in this context is the number of poses captured, which is a function of the variability of poses of the individuals, the number of frames of video included and the differences between selected frames: DeepLabCut specifically uses a k-means algorithm to facilitate selection of frames that capture variability across the videos (Nath et al. 2019; Mathis et al. 2018). The high accuracies we achieve with our training set size of thousands of frames indicates that the rate of change between usefully different poses across frames must be reasonably high, and that our sample sizes in the low thousands of frames are adequate to capture most of the variability in the range of landmark appearance. A sample size in the low thousands is consistent with estimates of the number of samples typically considered required in other machine-learning contexts (Cho et al. 2015).

Additionally, in this current study, we deliberately started with a dataset with a large diversity of infants' poses, skin color, background linen and nappy colours and videoing conditions such as shadow contrast and viewing angle. This range reflects the true diversity of clinical recordings currently taken. Judicious selection of the training data to include in this variation may assist in the construction of a trained model that is robust to different video conditions from an efficiently sized training set. Deliberate engineering for robustness to skin colour is essential to avoid the introduction of machine-learning performance bias that has been shown to impact different communities when these issues are not considered (Singh et al. 2022; Daneshjou et al. 2021).

Limitations and future work

This work has a number of limitations that can be addressed in future studies. Firstly, the testing set size used here, three videos, is very constrained, and the algorithm should be validated on a large dataset of clinically recorded videos, including high-risk identified infants. Secondly, we cannot tell from our results how much better performance will get with additional training data, and testing with a larger sample size of clinical recordings is warranted. According to our Kalman-filter performance metric, which is beneficial for being able to be calculated automatically across a whole video, we are currently performing at nearly perfect performance; however, comparison of the tracked positions against human labels may provide a more sensitive measure of the precision of the labelling, albeit necessarily on a smaller sample. Thirdly, we have only examined performance of one deep-net model, ResNet152, using the DeepLabCut platform. There has been a recent proliferation of similar models for pose estimation, and the performance of these in the context of infant tracking should be evaluated. A tool for configuring, training and evaluating all of these algorithms across a common training/testing set would greatly facilitate this work and likely be useful in many other domains. Fourthly, to ensure equitable healthcare outcomes, it is essential for future work to formally assess potential biases in machine-learning models, such as those related to skin colour or variety of poses in the training set, and to aim at mitigating these biases while enhancing the model's performance across diverse populations.

Fifthly, there is no evidence in the literature on the quantification of infant distal segment rotations in the context of General Movements, and so it is not yet clear whether, or to what extent, these movement features inform us about neurodevelopmental function and health. The performance of classifiers trained to detect neurodevelopmental concerns from the infant kinematics derived with or without rotational information should indicate the relative importance of these movements as a clinical indicator. In the advent that these features are found to be important, it will be fascinating to explore the functional implications of these movements in terms of healthy early motor learning and neuromuscular development.

Conclusion

This validation study confirms feasibility of tracking in infants between two and five months of age, placing landmarks that capture rotational motion of distal limb segments using markerless motion tracking technology from standard 2D clinically recorded videos. The required increase in specificity of the labelling of distal landmarks results in a performance drop in the tracking, as assessed using an automated approach that captures all pose variability in the test videos. However, this performance drop can be mitigated by increasing the size of the dataset. Relatively modest datasets with fewer than 30 videos can be expected to capture the range of landmark variation seen from infants in this age group. These results indicate the suitability of this approach for inclusion in an automated clinical platform for infant GMA.

Acknowledgements

- The algorithm development, data analysis and manuscript writing/preparation were undertaken by H. A. Data was recorded by S.M. and L.L. Manuscript was reviewed and revised by A. M., T. B. and S. W. Funding acquisition: T. B. and A. M. The final submitted article has been revised and approved by all authors.
- We are grateful for the support of the New Zealand eScience Infrastructure (NeSI) high performance computing facilities for hosting our computational processing. URL: <https://www.nesi.org.nz>
- We would like to thank the Child Development Centre at Waikato DHB, New Zealand, in particular Mrs Nikki Laker and Karli Joll, for their involvement in this study, data collection, preparation and providing access to the clinical recordings.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research was supported by Friedlander Foundation grant (Grant: 3720759). SW is funded by the Aotearoa Foundation.

ORCID

H. Abbasi  <http://orcid.org/0000-0003-1136-3280>

References

- Abbasi H, Battin MR, Butler R, et al. 2023. Early signatures of brain injury in the preterm neonatal EEG. *Signals*. 4(3):630–643.
- Abbasi H, Mollet SR, Williams SA, et al. 2023. Deep-learning for automated markerless tracking of infants general movements. *International Journal of Information Technology*. 1–11. <https://link.springer.com/article/10.1007/s41870-023-01497-z#citeas>.
- Ahearne CE, Boylan GB, Murray DM. 2016. Short and long term prognosis in perinatal asphyxia: an update. *World Journal of Clinical Pediatrics*. 5:67. doi:10.5409/wjcp.v5.i1.67.
- Bernava GM, Leo M, Carcagni P, Distante C. 2022. An advanced tool for semi-automatic annotation for early screening of neurodevelopmental disorders. *Image Analysis and Processing, ICIAP 2022 Workshops: ICIAP International Workshops, Lecce, Italy, May 23–27, 2022, Revised Selected Papers, Part II*. p.154–164.
- Bosanquet M, Copeland L, Ware R, Boyd R. 2013. A systematic review of tests to predict cerebral palsy in young children. *Developmental Medicine & Child Neurology*. 55:418–426. doi:10.1111/dmcn.12140.
- Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y. 2019. Openpose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 43:172–186. doi:10.1109/TPAMI.2019.2929257.
- Caruso A, Gila L, Fulceri F, et al. 2020. Early motor development predicts clinical outcomes of siblings at high-risk for autism: insight from an innovative motion-tracking technology. *Brain Sciences*. 10:379. doi:10.3390/brainsci10060379.
- Cho J, Lee K, Shin E, Choy G, Do S. 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. *arXiv preprint arXiv:1511.06348*.
- Cunningham R, Sánchez MB, Butler PB, Southgate MJ, Loram ID. 2019. Fully automated image-based estimation of postural point-features in children with cerebral palsy using deep learning. *Royal Society Open Science*. 6:191011. doi:10.1098/rsos.191011.
- Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. 2021. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatology*. 157:1362–1369. doi:10.1001/jamadermatol.2021.3129.
- Darsaklis V, Snider LM, Majnemer A, Mazer B. 2011. Predictive validity of Prechtl's method on the qualitative assessment of general movements: a systematic review of the evidence. *Developmental Medicine & Child Neurology*. 53:896–906. doi:10.1111/j.1469-8749.2011.04017.x.
- Einspieler C, Prechtl HF. 2005. Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. *Mental Retardation and Developmental Disabilities Research Reviews*. 11:61–67. doi:10.1002/mrdd.20051.
- Fairhurst C. 2012. Cerebral palsy: the whys and hows. *Archives of Disease in Childhood-Education and Practice*. 97:122–131. doi:10.1136/edpract-2011-300593.
- Ferrari F, Einspieler C, Prechtl H, Bos AF, Cioni G. 2004. Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants. London: Mac Keith Press. p. 1–104.
- Garcia JM, Gherpelli JLD, Leone CR. 2004. The role of spontaneous general movement assessment in the neurological outcome of cerebral lesions in preterm infants. *Jornal de pediatria*. 80:296–304. doi:10.2223/1203.
- Groos D, Adde L, Aubert S, et al. 2022a. Development and validation of a deep learning method to predict cerebral palsy from spontaneous movements in infants at high risk. *JAMA Network Open*. 5:e2221325. doi:10.1001/jamanetworkopen.2022.21325.
- Groos D, Adde L, Støen R, Ramampiaro H, Ihlen EA. 2022b. Towards human-level performance on automatic pose estimation of infant spontaneous movements. *Computerized Medical Imaging and Graphics*. 95:102012. doi:10.1016/j.compmedimag.2021.102012.
- Gunn AJ, Thoresen M. 2015. Animal studies of neonatal hypothermic neuroprotection have translated well in to practice. *Resuscitation*. 97:88–90. doi:10.1016/j.resuscitation.2015.03.026.
- Hadders-Algra M. 2014. Early diagnosis and early intervention in cerebral palsy. *Frontiers in Neurology*. 5:185. doi:10.3389/fneur.2014.00185
- Hadders-Algra M. 2018. Neural substrate and clinical significance of general movements: an update. *Developmental Medicine & Child Neurology*. 60:39–46. doi:10.1111/dmcn.13540

- Image-net challenge. (n.d.). <https://image-net.org/challenges/LSVRC/>.
- Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 25:1097–1105.
- Krizhevsky A, Sutskever I, Hinton GE. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*. 60:84–90. doi:10.1145/3065386.
- Labuguen R, Matsumoto J, Negrete S, et al. 2021. Macaquepose: a novel “in the wild” macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioral Neuroscience*. 14:581154. doi:10.3389/fnbeh.2020.581154.
- Lempereur M, Rousseau F, Rémy-Néris O, et al. 2020. A new deep learning-based method for the detection of gait events in children with gait disorders: proof-of-concept and concurrent validity. *Journal of Biomechanics*. 98:109490. doi:10.1016/j.jbiomech.2019.109490.
- Liu X, Yu S, Flierman NA, et al. 2021. Optiflex: multi-frame animal pose estimation combining deep learning with optical flow. *Frontiers in Cellular Neuroscience*. 15:621252. doi:10.3389/fncel.2021.621252.
- Mathis A, Biasi T, Schneider S, et al. 2021. Pretraining boosts out-of-domain robustness for pose estimation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. p. 1859–1868.
- Mathis A, Mamidanna P, Cury KM, et al. 2018. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*. 21:1281–1289. doi:10.1038/s41593-018-0209-y.
- Morgan C, Romeo DM, Chorna O, et al. 2019. The pooled diagnostic accuracy of neuroimaging, general movements, and neurological examination for diagnosing cerebral palsy early in high-risk infants: a case control study. *Journal of Clinical Medicine*. 8:1879. doi:10.3390/jcm8111879.
- Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW. 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*. 14:2152–2176. doi:10.1038/s41596-019-0176-0.
- NeSI. (n.d.). New Zealand eScience Infrastructure (NeSI), New Zealand.
- Prechtl HF. 1990. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development*. 23(3):151–158. doi:10.1016/0378-3782(90)90011-7.
- Prechtl HF. 1997. State of the art of a new functional assessment of the young nervous system. An early predictor of cerebral palsy. *Early Human Development*. 50:1–11. doi:10.1016/S0378-3782(97)00088-1.
- Raghuram K, Orlandi S, Church P, Luther M, Kiss A, Shah V. 2022. Automated movement analysis to predict cerebral palsy in very preterm infants: an ambispective cohort study. *Children*. 9:843. doi:10.3390/children9060843.
- Sakkos D, Mccay KD, Marcroft C, Embleton ND, Chattopadhyay S, Ho ES. 2021. Identification of abnormal movements in infants: a deep neural network for body part-based prediction of cerebral palsy. *IEEE Access*. 9:94281–94292. doi:10.1109/ACCESS.2021.3093469.
- Shin HI, Shin H, Bang MS, et al. 2022. Deep learning-based quantitative analyses of spontaneous movements and their association with early neurological development in preterm infants. *Scientific Reports*. 12:3138. doi:10.1038/s41598-022-07139-x.
- Silva N, Zhang D, Kulvicius T, et al. 2021. The future of general movement assessment: the role of computer vision and machine learning—a scoping review. *Research in Developmental Disabilities*. 110:103854. doi:10.1016/j.ridd.2021.103854.
- Singh R, Majumdar P, Mittal S, Vatsa M. 2022. Anatomizing bias in facial analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*. p. 12351–1212358.
- Spittle AJ, Boyd RN, Inder TE, Doyle LW. 2009. Predicting motor development in very preterm infants at 12 months’ corrected age: the role of qualitative magnetic resonance imaging and general movements assessments. *Pediatrics*. 123:512–517. doi:10.1542/peds.2008-0590.
- Wei K, Kording KP. 2018. Behavioral tracking gets real. *Nature Neuroscience*. 21:1146–1147. doi:10.1038/s41593-018-0215-0.
- Wu Q, Xu G, Wei F, et al. 2022. Supine infant pose estimation via single depth image. *IEEE Transactions on Instrumentation and Measurement*. 71:1–11.