PLOS DIGITAL HEALTH

RESEARCH ARTICLE

# Automated identification of abnormal infant movements from smart phone videos

**E. Passmore[1,2,3,4]*, A. L. Kwong[3,5,6], S. Greenstein[1], J. E. Olsen[5,6], A. L. Eeles[5,6], J. L. Y. Cheong[3,5,6], A. J. Spittle[3,5], G. Ball[1,3]**

**1** Murdoch Children's Research Institute, Developmental Imaging, Melbourne, Australia, **2** University of Melbourne, Engineering and Information Technology, Melbourne, Australia, **3** University of Melbourne, Medicine, Dentistry & Health Sciences, Melbourne, Australia, **4** Royal Children's Hospital, Gait Analysis Laboratory, Melbourne, Australia, **5** Murdoch Children's Research Institute, Victorian Infant Brain Studies, Melbourne, Australia, **6** Royal Women's Hospital, Newborn Research Centre, Melbourne, Australia

* elyse.passmore@mcri.edu.au





**Citation:** Passmore E, Kwong AL, Greenstein S, Olsen JE, Eeles AL, Cheong JLY, et al. (2024) Automated identification of abnormal infant movements from smart phone videos. PLOS Digit Health 3(2): e0000432. https://doi.org/10.1371/journal.pdig.0000432

**Editor:** Danilo Pani, University of Cagliari: Universita degli Studi Di Cagliari, ITALY

**Received:** August 31, 2023

**Accepted:** December 17, 2023

**Published:** February 22, 2024



**Data Availability Statement:** Data availability: Requests for data that supports the findings of this study can be made to the Murdoch Children's Research Institute Data Office (data.requests@mcri.edu.au). The data is not publicly available due to privacy and ethical restrictions. Code availability: We used DeepLabCut to track body points in videos. Model for body point tracking, code for pre-processing body point data after DeepLabCut, training machine-learning models, analysis of the results are available in our

## Abstract

Cerebral palsy (CP) is the most common cause of physical disability during childhood, occurring at a rate of 2.1 per 1000 live births. Early diagnosis is key to improving functional outcomes for children with CP. The General Movements (GMs) Assessment has high predictive validity for the detection of CP and is routinely used in high-risk infants but only 50% of infants with CP have overt risk factors when they are born. The implementation of CP screening programs represents an important endeavour, but feasibility is limited by access to trained GMs assessors. To facilitate progress towards this goal, we report a deep-learning framework for automating the GMs Assessment. We acquired 503 videos captured by parents and caregivers at home of infants aged between 12- and 18-weeks term-corrected age using a dedicated smartphone app. Using a deep learning algorithm, we automatically labelled and tracked 18 key body points in each video. We designed a custom pipeline to adjust for camera movement and infant size and trained a second machine learning algorithm to predict GMs classification from body point movement. Our automated body point labelling approach achieved human-level accuracy (mean ± SD error of 3.7 ± 5.2% of infant length) compared to gold-standard human annotation. Using body point tracking data, our prediction model achieved a cross-validated area under the curve (mean ± S.D.) of 0.80 ± 0.08 in unseen test data for predicting expert GMs classification with a sensitivity of 76% ± 15% for abnormal GMs and a negative predictive value of 94% ± 3%. This work highlights the potential for automated GMs screening programs to detect abnormal movements in infants as early as three months term-corrected age using digital technologies.

## Author summary

Cerebral palsy (CP) is the most common cause of physical disability in childhood and is a permanent lifelong condition. To improve functional outcomes, early diagnosis is crucial to facilitate early intervention. The General Movements (GMs) Assessment has high predictive validity for the detection of CP and is routinely used in high-risk infants. However,

GitHub repository https://github.com/epassmore/infant-movements.

**Funding:** This work was supported by the Murdoch Children's Research Institute (Clinician Scientist Fellowship to EP), Rebecca L Cooper Medical Research Foundation (PG2019421 to GB), National Health and Medical Research Council Investigator Grant (1194497 to GB; 2016390 to JC), NVIDIA Corporation Hardware Grant program and The Royal Children's Hospital Foundation, Melbourne. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** Alicia Spittle is a tutor with the General Movements Trust. All other authors have no conflicts of interest to declare.

only 50% of infants with CP have overt risk factors at birth. There is a need for CP screening programs, but access to trained assessors limits their feasibility. To solve this problem, we developed a deep learning method for automating GMs assessment from smart phone videos. We acquired 503 videos captured at home by parents of infants aged 12 to 18 weeks term corrected. Our deep learning algorithm tracked 18 key body points on the infant throughout the video with human-level labelling accuracy. A custom data processing pipeline was designed to account for infant size, camera movements and different video formats. A second machine learning algorithm was trained to predict GMs classification and validated in a cohort of preterm (high-risk) and term-born infants. Our work highlights how digital technologies can be used to automate GMs assessment, allowing infants to be screened for abnormal movements as early as three months old at home.

## Introduction

Cerebral palsy (CP) refers to a group of disorders that affect motor development, movement and posture and are attributed to non-progressive disturbances or injuries to the developing brain before 1 year of age [1]. Cerebral palsy is the most common cause of physical disability during childhood, occurring at a rate of 2.1 per 1000 live births worldwide [2]. While those born preterm or with low birthweight are at greater risk of having CP, almost 50% of infants with CP are born at term without overt risk factors [3,4].

Early diagnosis is essential to improve clinical and functional outcomes of children with CP [5]. Detecting abnormal motor development within the first 6 months after birth allows targeted interventions, coincident with periods of rapid neurodevelopmental plasticity and musculoskeletal development. It has been shown that early intervention improves children's motor and cognitive development as well parental wellbeing [5,6]. However, the average age of CP diagnosis is 19 months [3], and only 21% of infants with CP are diagnosed before 6 months [3,7], thus many infants miss a crucial window for early intervention.

The General Movements (GMs) Assessment can accurately predict those at highest risk of CP within the first few months after birth [8,9]. General movements are spontaneous movements involving the whole body with a changing sequence of arm, legs neck, and trunk movements [10]. Between nine and twenty weeks of age, spontaneous movements are characterised by continuous small movements with moderate speed and variable acceleration, termed 'fidgety' movements [11]. These 'fidgety movements' are typically recognised using a trained assessor's gestalt perception, while the infant is lying awake on their back with no direct handling or interaction [11]. This assessment is best completed from video recordings of the infant and has high predictive validity for neurodevelopmental outcomes and excellent inter-rater reliability [9,12,13]. The GMs assessment when used during the 'fidgety period' has the potential to be an important screening tool in the diagnosis of CP.

The specialized training required by GMs assessors means that many primary care services and hospitals do not offer routine GMs assessment, which limits the widespread adoption of GMs assessment as a screening tool [14]. The ability to assess GMs using video recordings has raised the potential to improve equitable healthcare access for those living in remote regions or in low-resource settings. Recently, the development of smartphone apps that allow the standardised recording of video by an infant's primary carers using a hand-held device, have been shown to improve access to GMs assessment and allow identification of high-risk infants outside of clinical settings [12,13,15,16]. Automated scoring of GMs from video can provide a

mechanism to overcome these bottlenecks and enable high-throughput assessment for screening programs.

Recent advantages in computer vision and deep learning have led to the emergence of pose estimation techniques, a class of algorithms designed to estimate the spatial location of a set of body points from videos or pictures, and track movements with high accuracy [17–20]. Pose estimation tools do not require any specialised equipment, movement sensors or markers, and can be readily applied to track movement in new videos once trained. Several open-source pose estimation tools, pre-trained on large databases of human movement, are available for direct application to new datasets [17,21–23]. However, the standard implementation of such algorithms has been found to perform poorly in videos of infants, likely due to significant differences in body segment size and scale [24,25]. Thus, fine-tuning or re-training of pose estimation models is required to accommodate the anatomical proportions of infants [24,26,27]. Further, videos acquired outside of controlled, clinical or research laboratory settings may vary significantly in terms of camera angle, length, resolution, and distance to subject, requiring additional processing steps before analysis [24,28].

Several recent studies have yielded promising results predicting motor outcomes in infants using movement tracking data from pose estimation tools [24,28–32]. Using a semi-automated approach with manual key point annotation of clinical videos, Ihlen *et al*. demonstrated computer-based movement assessments can perform comparably with observation-based GMs in predicting CP (area-under-ROC-curve, (AUC) = 0.87) [31]. Using videos acquired in a specialised laboratory setting and an infant-adapted OpenPose model, Chambers *et al*. employed a Bayesian model of joint kinematics to identify infants at high-risk for motor impairment [24]. Recent applications of deep learning models to classify movement data have also reported good performance, with one example detecting the presence or absence of fidgety movements in 5-second video clips with 88% accuracy in a laboratory setting (n = 45 infants) [30]. In a large, multisite study of high-risk infants (15% with CP) Groos et al. reported a positive predictive value of 68% (negative predictive value of 95%) for later CP diagnosis using an ensemble of Graph Convolutional Networks (GCN) applied to clinical videos [29]. Similarly, Nguyen-Thai et al. applied GCNs to OpenPose tracking data to create a spatiotemporal model of infant joint movement in 235 smartphone videos, reporting an average AUC of 0.81 for the prediction of abnormal fidgety movements [28].

Despite initial progress, significant challenges remain for the application and uptake of this technology. Many studies to-date have been limited by small sample size (typically $< 100$ infants) and few have been conducted outside of clinical or laboratory settings [32–34]. In this study, using a large cohort of infant movement videos (n = 503) captured remotely via a dedicated smart phone app, we test the efficacy of automatic pose estimation and movement classification using deep learning methods to predict GMs classification. In addition, we design a custom processing pipeline to accommodate video capture from hand-held devices, identify factors that adversely affect automatic body point labelling accuracy and locate salient movement features that predict abnormal outcomes in individuals.

## Results

### Automated body point labelling with human-level accuracy

We acquired 503 3-minute videos from 341 infants at 12 to 18 weeks' term-corrected age using Baby Moves, a dedicated smartphone app [12]. To fine-tune a pose estimation model for infant videos, we created a training dataset using a diverse selection of n = 500 frames from 100 videos (5 frames per video, see **Methods**). We manually annotated eighteen body points (crown, eyes, chin, shoulders, elbows, wrists, hips, knees, heels and toes; **Figure A in S1 Appendix**)

and trained a fully-customisable pose estimation model, Deep Lab Cut (DLC) [35] to automatically detect each body point (Fig 1). Body point labelling using the trained DLC model was highly accurate (Fig 1E) achieving a mean difference between manual and automatic annotations of 3.7% of infant length (SD: 5.2). DLC performance was comparable to inter-rater reliability (IRR) of two annotators (average ± SD: 3.6 ± 3.7% of infant length; Fig 1E). Labelling accuracy varied moderately across body points, with the highest accuracy for the eyes (average ± SD: manual/auto 1.6 ± 1.3% infant length; IRR 1.0 ± 0.5% infant length) and lowest accuracy for the hips (average ± SD: manual/auto 6.3 ± 5.1% infant length; IRR 7.2 ± 3.1% infant length) (**Figure B, Table A in** S1 Appendix). There was no significant difference in labelling performance between video resolutions (**Table B in** S1 Appendix).

During labelling, the DLC model assigned each point a measure of prediction confidence. After removing points labelled with low confidence (see Methods), we found that the percentage of frames labelled on average was 92% (SD: 16%). The percentage of frames in which each point was confidently labelled was lowest for the wrists (average ± SD Left: 81 ± 21%, Right 78 ± 24%) and heels (average ± SD: Left: 69 ± 24%, Right 77 ± 20%) (**Figure C in** S1 Appendix), due in part to these body points being occluded by other body parts at instances throughout the video and exhibiting a greater range of movement. We conducted a sensitivity analysis to determine potential factors that related to labelling failures (see Methods). We found that the amount of clothing worn by the infant moderately affected model performance with outfits that covered the hands and feet adversely affecting labelling of the extremities ($F = 5.180$, $p = 0.006$; **Table C in** S1 Appendix). As a quality control step, only videos where on average at least 70% of body points per frame were confidently labelled were included for further analysis. After quality control, our final cohort comprised n = 484 videos from 327 unique infants.

## Predicting GMs from video data

As videos were not acquired in standardised clinical or experimental settings, positional data were pre-processed using a custom pipeline to account for different video acquisition parameters and potential camera movements relative to the subject, prior to classification. This consisted of body point mislabelling removal, gap filling, adjusting for camera movement, scaling to unit length based on infant size and framerate normalisation (see Methods). After pre-processing, each video was represented as a 46 × 4500 feature-by-frame matrix comprising standardized $x$ and $y$ coordinates of each body point and 2D joint angles of 10 joints in each frame.

Fidgety general movements may occur at different timepoints throughout the video occurring with different frequencies and movements, therefore we aimed to identify short periods where discriminant movements were present using a sliding window approach (Fig 2). We trained a convolutional neural network to predict GMs classification based on short instances (approximately 5-seconds) of positional data over time (Fig 2C), calculating video-level predictions by integrating over all clips for a given video. The classification model was trained to predict an outcome of either normal or abnormal GMs. For the purposes of our model infant videos classified as either abnormal or absent GMs were combined to form the abnormal category.

Averaged over 25 cross-validation repeats (70% train/15% validation/15% test), the trained model achieved an AUC (mean ± S.D.) of 0.795 ± 0.080 in unseen test data (Fig 2D) and balanced accuracy of 0.703 ± 0.083 (Fig 2E). For abnormal/absent GMs, the positive predictive value (PPV) was 0.277 ± 0.077 and the negative predictive value (NPV) was 0.941 ± 0.035. Sensitivity and specificity were 0.755 ± 0.150 and 0.651 ± 0.078, respectively (Fig 2E). Each video was included in the held-out test set on average 4 (SD: 2) times across the 25 folds (**Figure D in** S1 Appendix).

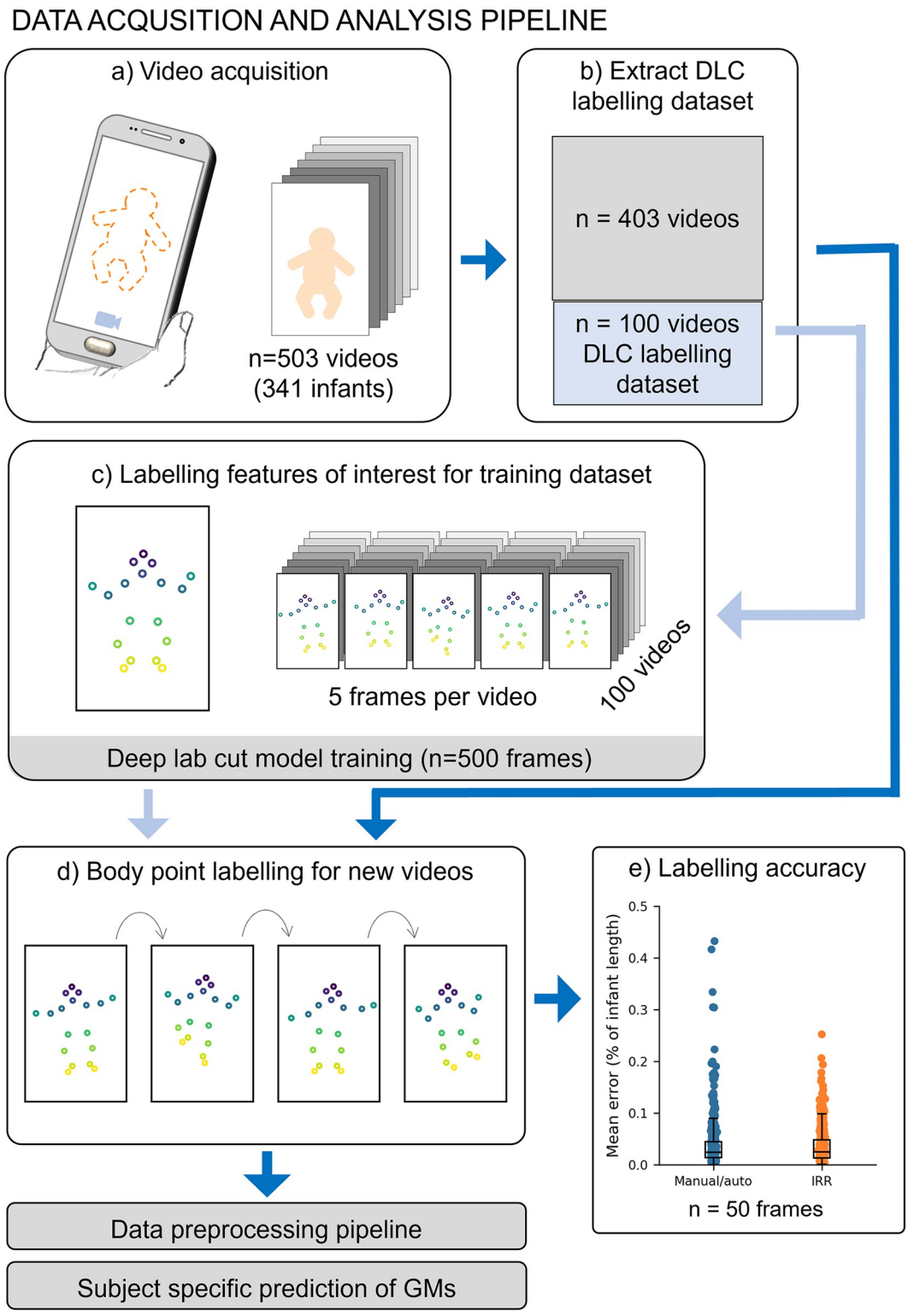


**Fig 1. Data acquisition and analysis pipeline. A.** Acquisition of 503 videos using the dedicated Baby Moves smartphone app [12]. **B.** 100 videos were selected for DLC training, stratified by age at video acquisition, sex and GMs classification. **C.** From each of the 100 training videos, five frames were selected for manual labelling using a k-means clustering algorithm (see **Methods**; total DLC training dataset: 500 frames). **D.** The trained DLC model was used to label all frames in all videos. This constitutes the full dataset with body point positional data used for GMs classification **E.** Labelling accuracy was evaluated in a subset of 50 frames

not included in the training dataset. The difference between manual and automatic labelling (Manual/auto) and inter-rater reliability (IRR) of two annotators was calculated and expressed as percentage of infant length.

https://doi.org/10.1371/journal.pdig.0000432.g001

Performance was consistent over a range of model parameters, including batch size, learning rate and weight regularisation (**Figure E in** S1 Appendix). The inclusion of video meta-data, age at video acquisition and birth cohort (extremely preterm or term-born infants) improved model performance significantly (**Figure E in** S1 Appendix), compared with classification using video data alone (AUC = 0.749 ± 0.077). Taking advantage of multiple model instances trained across different cross-validation repeats, we found that GMs prediction were generally consistent across different data splits in the cross validation (Fig 2D–2E, Fig 3).

We compared performance with an alternative baseline model: an $l_2$-regularised logistic regression applied to a set of timeseries features extracted from each video [36]. The baseline model achieved an AUC of 0.706 ± 0.098 (0.604 ± 0.106 without video meta-data). A nonlinear, kernelized logistic regression model achieved cross-validated AUC = 0.720 ± 0.100.

## Examining spatial and temporal model attention during prediction

To identify potential features that were important to model prediction, we computed spatio-temporal saliency maps for each video [37], (Fig 4). This value highlights features (for a given body point in a single video frame) where changes in input would elicit the largest change in model prediction and can be used as a measure of model sensitivity to input data [37,38]. An example saliency map is shown for a single subject in Fig 4. Saliency varied across the length of the video, corresponding with variations in model attention (white line, Fig 4). Clips with high saliency, relative to all subjects in the test set, are highlighted with yellow bars on the input feature timeseries, illustrating model attention to periods of different length spread throughout the video. Averaging total saliency across all clips for each body point reveals higher model sensitivity to position of the lower body points (Fig 4 middle), including movement of the knee and ankle joints.

Similar patterns of model saliency were observed across all participants. A map of group average feature saliency (averaged across clips, participants and cross-validation repeats) is shown in Fig 5A. Model saliency was highest in the lower body. This pattern was consistent across cross-validation repeats (**Figure F in** S1 Appendix) and between normal and abnormal/absent GMs prediction (**Figure G in** S1 Appendix).

To further characterise features to which the model prediction was sensitive, we compared timeseries data in clips with high (90$^{th}$ percentile) and low (10$^{th}$ percentile) total saliency (Fig 5B–5C). The number of high saliency clips did not differ between normal (mean ± S.D. = 55.21 ± 33.93) and abnormal/absent (51.74 ± 35.56) GMs videos (**Figure H in** S1 Appendix). For each clip, we calculated the mean (absolute) displacement of body points from the average position, as well as the standard deviation of displacements over frames. We found that, high saliency clips were characterised by body point positions closer to the average body position (Fig 5B) and by a lower standard deviation of joint displacements over time compared to clips with low saliency (Fig 5C).

## GMs prediction and development at 2 years

We compared our model GMs predictions with participant's motor, cognitive and language outcomes at 2-years corrected age as assessed by the Bayley Scales of Infant and Toddler Development-3$^{rd}$ edition (Bayley-III) (Fig 6; **Figure I, Figure J in** S1 Appendix). We found strong evidence for differences in 2-year motor composite scores between infants with different

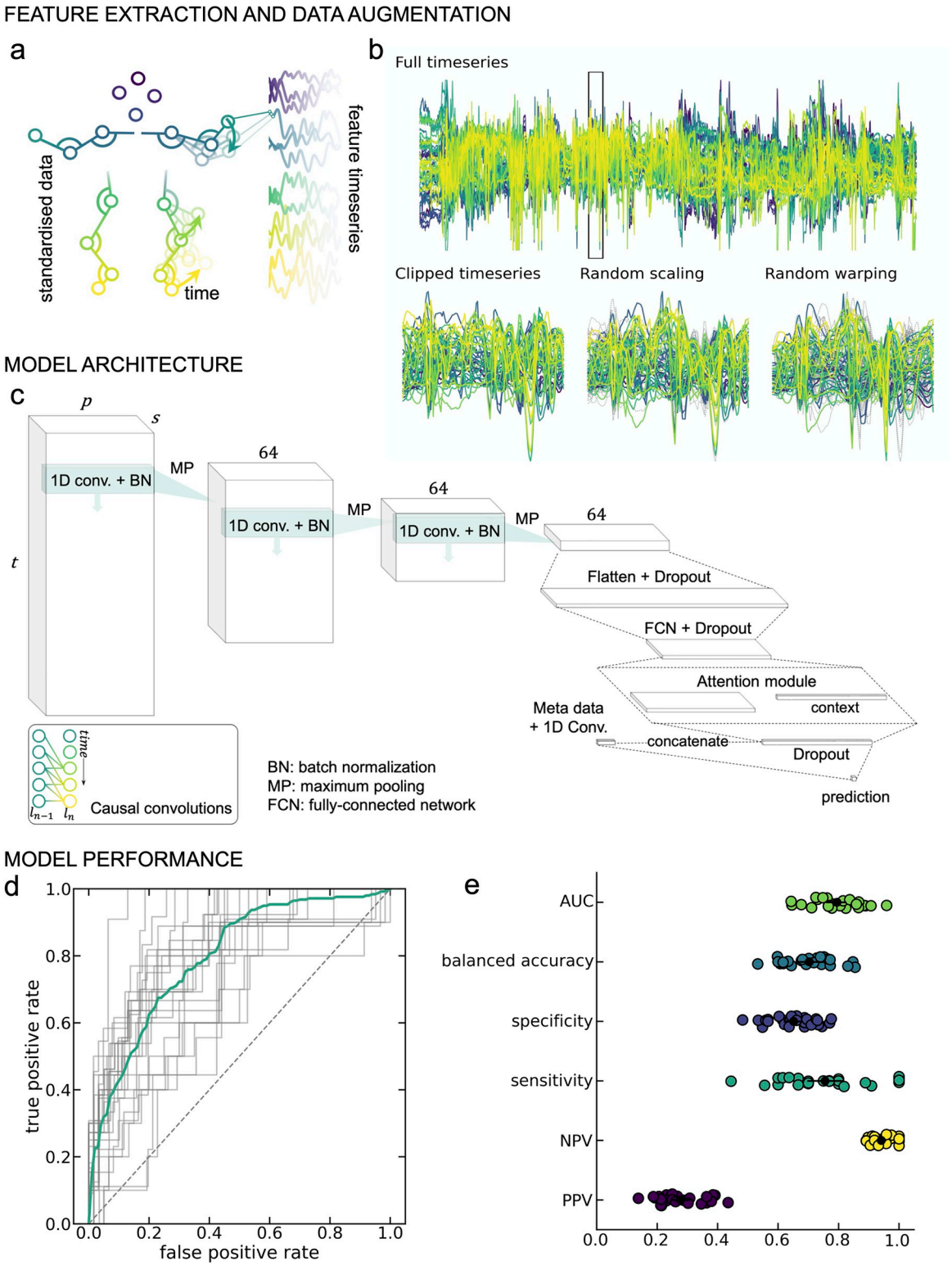


**Fig 2. GMs prediction from movement data. A**. Framewise positional data from DLC labelled videos were preprocessed to derive a set of feature timeseries (46 features × 4500 frames) per video. **B.** The classification model was trained on 128-frame clips for the full timeseries (top). Data augmentation steps (magnitude scaling and time warping; bottom middle and right) were applied to each clip during training. For each augmentation method, dashed grey lines indicate timeseries position prior to the augmentation step. **C.** Model architecture. 1D convolutional layers were combined with an attention module to classify GMs. Causal convolutions (inset) were applied

to account for the temporal structure of the data. **D**. Receiver-operator curves (ROC) for each of the 25 cross-validation repeats. The mean curve is overlaid in teal. **E**. Model performance statistics for each of the cross-validation repeats. Mean and standard deviation across repeats are overlaid in black. AUC = area under the ROC; NPV = negative predictive value; PPV = positive predictive value.

https://doi.org/10.1371/journal.pdig.0000432.g002

predicted GMs classifications (normal vs abnormal\absent) when metadata (age at acquisition and birth cohort; extremely preterm or term-born infants) were included in the model, mean difference 10.70 (95%CI = [6.77, 14.62], t(255) = 5.368, p<0.001). These differences were diminished when using movement data alone, mean difference 1.74 (95%CI = [-2.52, 5.99], t (255) = 0.805, p = 0.421) (Fig 6A). There was also strong evidence for differences in 2-year cognition and language composite scores between predicted GMs classifications (**Figure I, Table D in** S1 Appendix).

## Discussion

Using deep learning applied to smart phone videos, we tracked infant movements at 12–18 weeks term-corrected age, predicting GMs classification outside of a controlled clinical setting. Our paper illustrates the potential for early automated detection of abnormal infant movements implemented through at-home video acquisition.

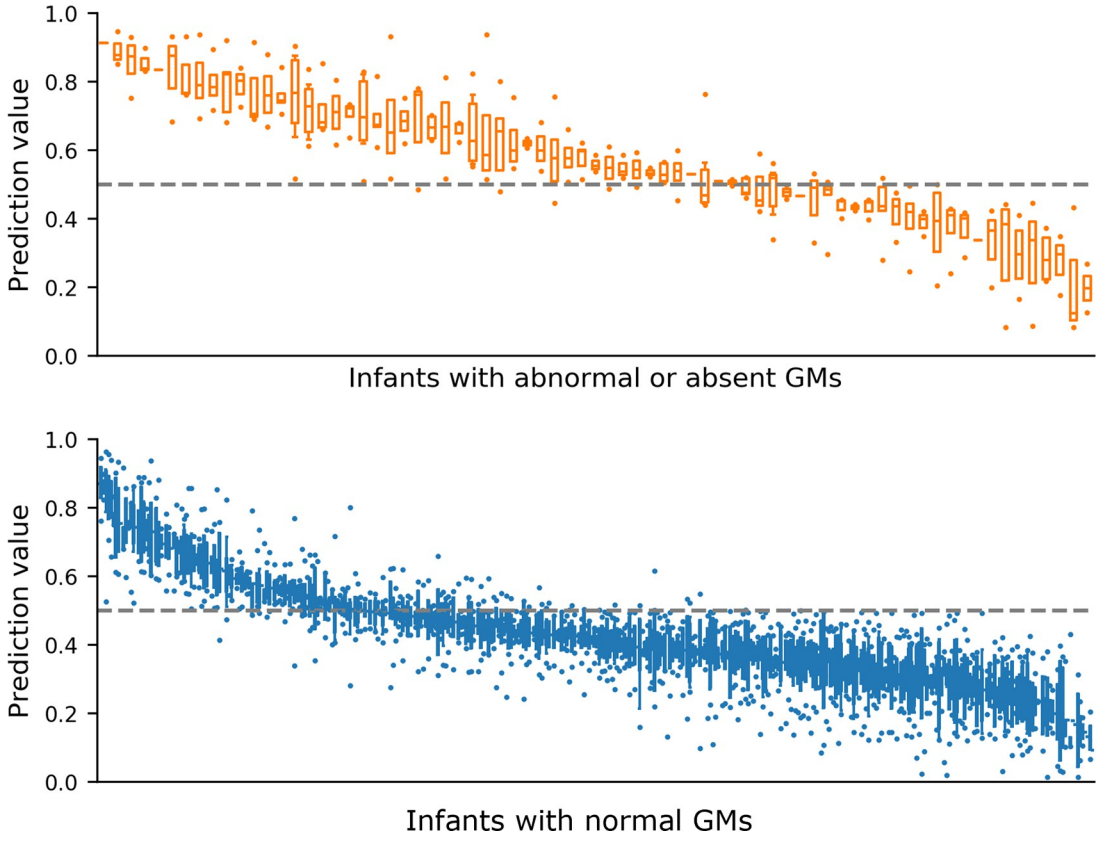


**Fig 3. Variation in GMs prediction values from 25-fold cross validation.** Prediction values reported when infant movement included in held out test set, ordered by median prediction value. Top (orange) infant videos scored as abnormal or absent GMs by expert rater, n = 73 infant videos. Bottom (blue) infant videos scored as normal GMs by expert rater, n = 396 infant videos. Boxes with horizontal line represent interquartile range and median respectively, error bars represent 95% confidence interval and dots represent outliers. Dashed line at 0.5 represents cut-off value for classifier between prediction categories.

https://doi.org/10.1371/journal.pdig.0000432.g003

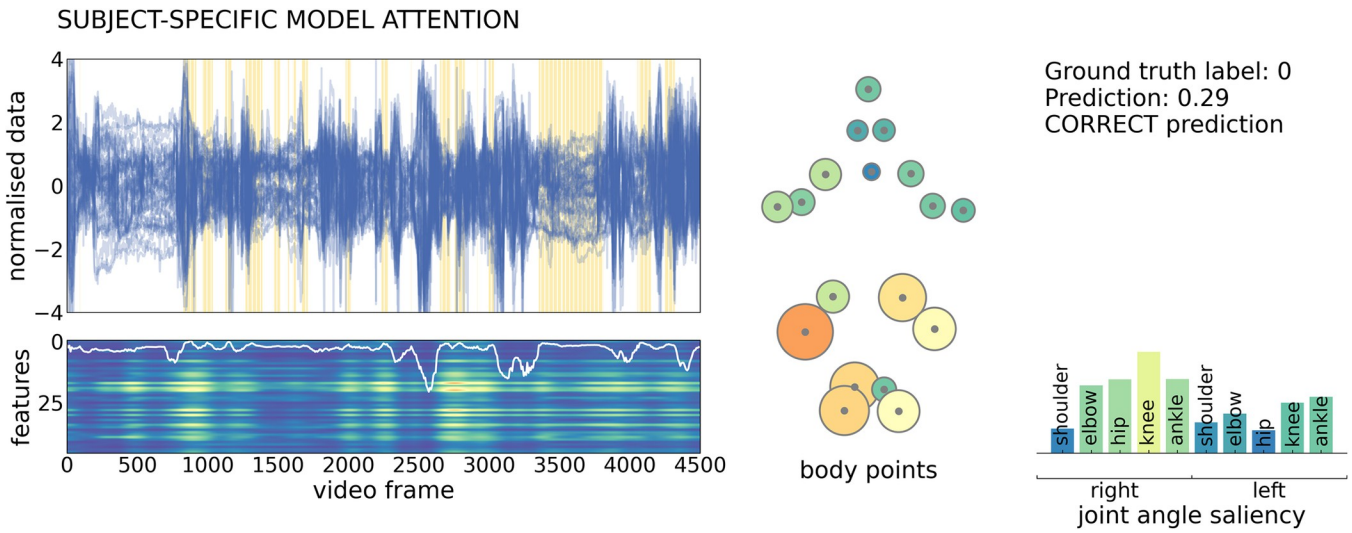


**Fig 4. Example of subject specific model attention to input features.** Feature timeseries (top) and saliency map (bottom) for a single, correctly-classified video from an infant with normal GMs. Timeseries are shown for each feature (n = 46), across the length of the video. Yellow bars indicate clips with high model attention (75$^{th}$ percentile across all subjects). Saliency was calculated for each feature in each frame and summed over frames within each clip (n = 547 clips). The map has been upsampled and smoothed to match the length of the timeseries (frames = 4500). Lighter colours indicate higher saliency (arbitrary unit). Clip attention derived from the attention module (upsampled and smoothed) is overlaid in white. Average saliency across the full video is shown for each body point (middle) and joint angle (right). Lighter colours and larger size reflect higher saliency. The model prediction is shown top right, where 0 indicates normal GMs prediction.

https://doi.org/10.1371/journal.pdig.0000432.g004

Our best performing model for predicting expert GMs classification, was a deep learning model, consisting of 1D convolutions and an attention module. Our model achieved an AUC 0.80 (SD: 0.08), comparable to results obtained from Ihlen et al. and Groos et al. which directly predicted CP in cohorts of high-risk infants using video recordings from stationary cameras in

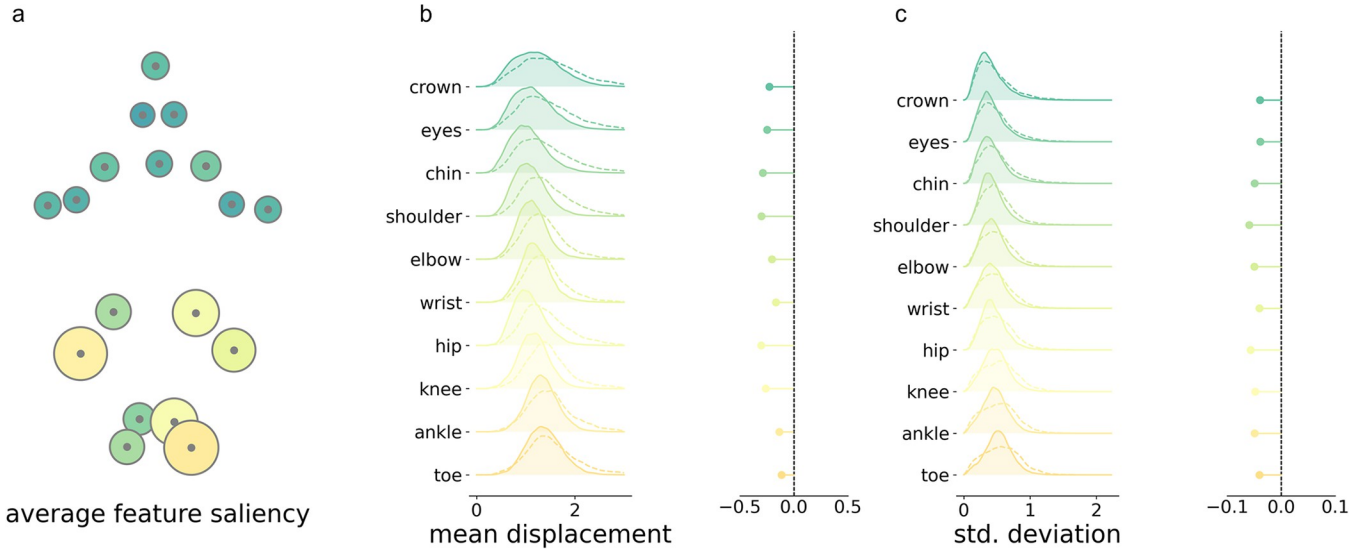


**Fig 5. Body point saliency for all infants. a**. Average saliency for all videos, lighter colours and larger size reflect higher saliency. **b.** Left, mean absolute distances between each body point and their respective average position during clips of high (solid line, filled) and low (dashed line) saliency. Density plots show the distribution of displacements for high and low clips over all videos and cross-validation repeats. Right, median difference between displacement in high and low clips. **c**. Left, standard deviation (std) of displacements from the average position in high and low saliency clips, over all videos and cross-validation repeats. Right, median difference in standard deviation distributions.

https://doi.org/10.1371/journal.pdig.0000432.g005

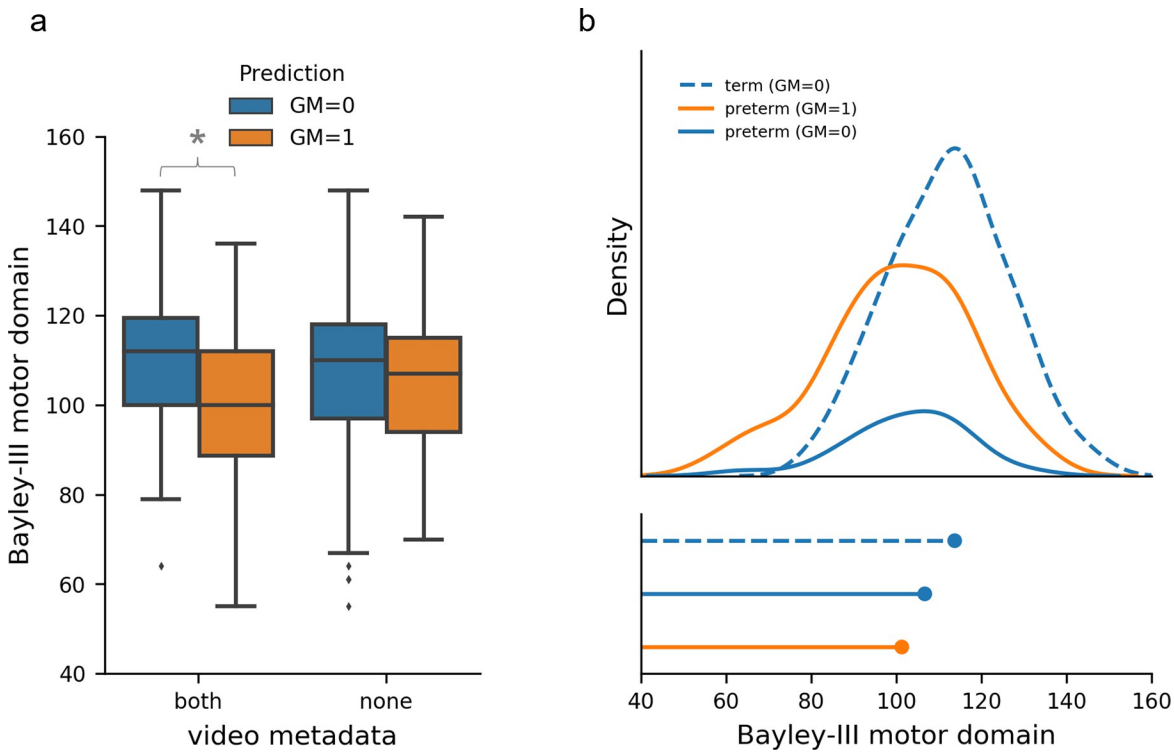


**Fig 6. GMs prediction and motor development at 2 years. A.** Bayley-III motor outcome stratified by GMs prediction (n = 252 infant videos) and model variants trained using video movement and metadata (both = age at acquisition and birth cohort) and movement data alone (none). Blue indicates GMs prediction = 0 (normal) and orange indicates GMs prediction = 1 (abnormal/absent) for all graphs. * Indicates strong evidence for differences between GMs classification prediction from independent two-sample t-test. **B.** Top, density function of motor outcome by birth cohort and GMs prediction. Bottom, Peak of density function. Term infants are represented by dashed lines and preterm infants by solid lines.

https://doi.org/10.1371/journal.pdig.0000432.g006

clinical settings [29,31]. Our model outperformed alternative baseline models and was robust over various hyperparameter settings. We demonstrated that including participant metadata is crucial to improving model predictions, highlighting the increased risk for abnormal movements in preterm born individuals [8]. Of the 41 infants with abnormal GMs as scored by trained assessors 35 (85%) were from preterm infants and 6 (15%) from term-born infants. We found that including metadata (birth cohort and age) improved model performance from an AUC = 0.70 based on movement data alone to AUC = 0.80. Notably, classifying on birth cohort alone would result in an AUC of 0.69, quantifying the added value of the movement data.

Our predicted GMs were associated with poorer neurodevelopmental outcome at 2 years of age. We found that this effect was largely dependent on preterm birth, although infants with predicted abnormal GMs scored lower on average (mean difference 10.70) regardless of birth cohort. However, most participants had Bayley's-III motor scores within published normative ranges (>85) [39]. The association between preterm birth and poor neurodevelopmental outcomes is well established [8,13] and this finding reflects the relatively lower predictive validity of GMs ratings, and therefore model predictions, for motor and cognitive outcomes at 2 years.

GMs classification is a strong predictor of CP [8,9]. We used abnormal or absent GMs as a surrogate measure for CP risk. Abnormal or absent fidgety GMs during this developmental window are both associated with neurodevelopmental impairment [8,13]. Combining abnormal and absent groups, who may have different movement signatures, into a single group may

have affected model performance but numbers were too small to further split the groups (n = 40 and 36 respectively). For the term-born infants that had absent GMs at the 12-week timepoint (n = 11) most of these normalised by 14-weeks (n = 8), which may have impacted our GMs prediction model's performance. An area for future research is training classification models that directly predict CP or other neurodevelopmental impairment [29,31]. We utilised a flexible framework for our GMs classifier, so that this is possible in the future. This was precluded in our dataset due to the reduced number of infants with two-year follow-up (252/371) and only 6 of those having a diagnosis of CP, while other neurodevelopmental impairments were not reported. To progress this area of research, access to large data sets that are diverse in both high- and low-risk infants and neurodevelopmental outcomes are required. Due to the identifiable nature of infant videos the sharing of data to form these large data sets is often challenging. While we are unable to provide the videos open access to progress research in this area, we have made our trained DLC infant pose-estimation model and GMs classification model openly available.

To track infant movement, we used a pose-estimation algorithm, Deep Lab Cut [35,40], which has the advantage of being customisable across species, age and features of interest using a minimal training dataset [40]. Our model achieved human-level labelling accuracy of body parts with a mean difference of 3.7% of infant length (SD: 5.2). Due to the non-controlled settings in which videos were acquired, we performed a sensitivity analysis, identifying video features that could affect labelling accuracy in at-home video recordings. Body point labelling was robust to background and video lighting but moderately affected by clothing worn by the infant, specifically clothing covering hands or feet. Use of at home video recordings introduced additional data processing challenges, including camera movement relative to the infant and different video formats (frame rate, resolution, distance from infant). To accommodate this, we developed a novel framework, that is versatile and supports videos taken outside a standardised clinical setting. There is increasing awareness around the inherent bias in pretrained pose-estimation models, with a bias towards adult males with lighter skin tone [41]. The dataset we used to train our pose-estimation model had an even distribution of sex however we lacked diversity in skin tone, in part due to study location in Melbourne Australia and the inclusion of only families that could speak English [12]. This is certainly an area where our pose-estimation model could be improved in the future. We choose to label the wrist and heel and big toe only in our pose-estimation model in part due to frame rate and resolution of our videos. However, more detailed anatomical annotations for the hands and feet maybe beneficial in recognition of the role those body parts play in identification of fidgety movement during GMs assessment [11].

We used 5 second clips to train our model, this allowed us to identify periods of movement that were informative to the model prediction. By analysing model saliency, we were able to extract information about which movement features were attended to by the model, discriminating between abnormal and normal GMs. We identified that lower limb movements contributed more to the classifier's output, with higher saliency attributed to video clips where infant position was closer to the average position, ignoring periods where the infant has moved significantly from the supine position (i.e.: out-of-frame movement, rolling). Other studies have used high-resolution video annotations identifying periods of abnormal movement within videos [31]. While this approach is likely to improve automated movement identification it requires a significant amount of manual annotation and labelling that would be difficult to achieve in larger cohorts.

Several recent studies have yielded promising results predicting motor outcomes in infants. However, to date these studies have been limited by small sample size (typically $< 100$ infants), only include high-risk infants and few have been conducted outside of clinical or laboratory

settings [32]. A strength of our study is the inclusion of both extremely preterm and term born infants within our dataset. It is essential that these automated approaches are validated on both preterm and term-born infants if they are to be implemented in population level screening programs. Our study offers an automated approach to assess GMs, that is capable of accommodating videos recording outside the clinical setting. Our work highlights the potential for automated approaches to screen for CP at a population level, which would enable increased access to early interventions for these children.

## Materials and methods

### Study design

This is a retrospective secondary analysis of data acquired as part of the prospective, multi-centre, population-based cohort study Baby Moves [12]. This study developed and evaluated a deep learning algorithm to track infant movement and predict GMs classification from infant movement videos (**Fig 1**). Written informed consent was obtained from parent/caregivers of infants prior to inclusion in the study. The study was approved by the Royal Children's Hospital Ethics Committee (HREC35237). Full details of the study protocol can be found in Spittle and colleagues [12].

### Participant data

Videos were recorded using the Baby Moves smart phone app by the parent/caregiver on their personal device between April 2016 and May 2017 [12]. The app prompted the parent/caregiver to record and upload two movement videos, one at 12-weeks and the second at 14-weeks term corrected age. Video upload was only possible in the app between 12-weeks and 17-weeks and 6 days term-corrected age. Videos were acquired from n = 155 (77 female [50%]) extremely preterm infants (<28 weeks' gestation) and 186 (91 female [49%]) term-born infants, (**Table E in S1 Appendix**). In total, 503 videos from n = 341 infants aged between 12- and 18-weeks term corrected age were available. For a subset of n = 160 (75 preterm, 85 term), two videos were collected per infant during this period.

### Video capture

To facilitate video recording outside of clinical or laboratory settings, the Baby Moves app provides detailed instructions and a dotted outline overlay to improve positioning of the infant in the video frame [12]. Guidance was given to parents/caregivers to perform the video while the infant was lying quietly and not fussing with minimal clothing, consisting of singlet and nappy only. Subsequently, videos were securely uploaded to a REDCap database [42,43] at the Murdoch Children's Research Institute for remote review. GMs were scored according to Prechtl's GMs Assessment [11] by two independent assessors that were unaware of participants' neonatal history. General movements were classified as normal if fidgety GMs were intermittently or continuously present, absent if fidgety GMs were not observed or sporadic, and abnormal if fidgety GMs were exaggerated in speed and amplitude. If there was disagreement between the two assessors, then a third experienced GMs trainer and assessor made the final decision. For the purposes of automating the GMs Assessment, videos scored as either absent or abnormal GMs were combined to form the abnormal group. Any videos rated as unscorable were not evaluated in this study.

All videos were submitted in MP4 format. Due to differences in device model and settings, three video resolutions were present in the dataset: 480 x 360 (n = 366 videos), 640 x 480 (13 videos) and 720 x 480 (126 videos) with a median frame rate of 30 frames per second (range:

15 to 31). Each video was 3 minutes in length resulting in mean (SD) 5100 (497) frames per video.

## Automated body point labelling of smart phone videos

We trained a deep learning model using Deep Lab Cut, version 2.1 (DLC) [35] to label and track key body points. To train the DLC model, we formed a training dataset consisting of a subset of 100 videos from our dataset stratified for age, sex, birth cohort (preterm or term) and video resolution. Only one video per infant was allowed in the training data set. For the training dataset, five frames from each video were manually labelled with 18 key body points: crown, chin, eyes, shoulders, elbows, wrists, hips, knees, heels and big toes (**Fig 1**; **Figure A in S1 Appendix**). Manual labelling was performed via the DLC graphical user interface. To ensure diversity of movements in the 5 frames selected for labelling, we used a k-means clustering algorithm implemented in DLC to select one frame from each of the five distinct clusters within each video for labelling. We implemented a DLC model with a pre-trained ResNet-50 backbone and trained for 1 million iterations on a NVIDIA TITAN Xp using a training/validation fraction of 0.95/0.05.

Once trained, the DLC model was used to automatically label the 18 body points for all videos in the dataset. For each frame, the DLC model returned the x- and y-coordinates in pixels of the body points relative to the corner of the video image and its prediction confidence. Body points with a prediction confidence below 0.2 were removed as recommended by Mathis et al. [35]. Labelling accuracy of the DLC model was evaluated on 50 frames, each from a different video not included in the training dataset. This evaluation dataset was stratified for age, sex, birth cohort (preterm or term) and video resolution to ensure diversity in images evaluated. DLC body point labelling was evaluated using the difference between manual and predicted labels for the evaluation dataset. To evaluate inter-rater reliability (IRR) for body point labelling a second human annotator repeated labelling on the same 50 frames. The difference between labels for the two annotators were calculated. To compare differences across different video resolutions body point labelling was expressed as a percentage of infant length (crown to mid hip).

Additional metrics of DLC model performance included the number of unlabelled body points per video. As videos were collected outside of a controlled clinical setting, we conducted a sensitivity analysis to determine whether variability in certain factors across the individual videos may influence model performance. Each video not included in the training dataset (n = 403) was categorised by the following factors: Lighting (Dark/Okay/Bright), clothing (Bodysuit/Nappy & singlet/Nappy only), skin tone (Light/Fair/Medium/Dark), infant in frame entire video (Yes/No), background (Pattern/Solid Colour–Dark/Solid Colour–Light), extra items in view (No/Another Child/Recorder's feet/Toys/Other). We tested if model performance was affected by the listed factors using a mixed model Analysis of Variance (ANOVA).

## Pre-processing pipeline

Data pre-processing consisted of quality control, outlier removal, gap filling, adjustment for camera movement, scaling and feature extraction (**Fig 7**).

## Quality control

To ensure high-quality movement data from each video was available for further analysis, we established a quality control measure based on body point labelling. Only videos in which more than 70% of body points were labelled on average across all frames were used. This

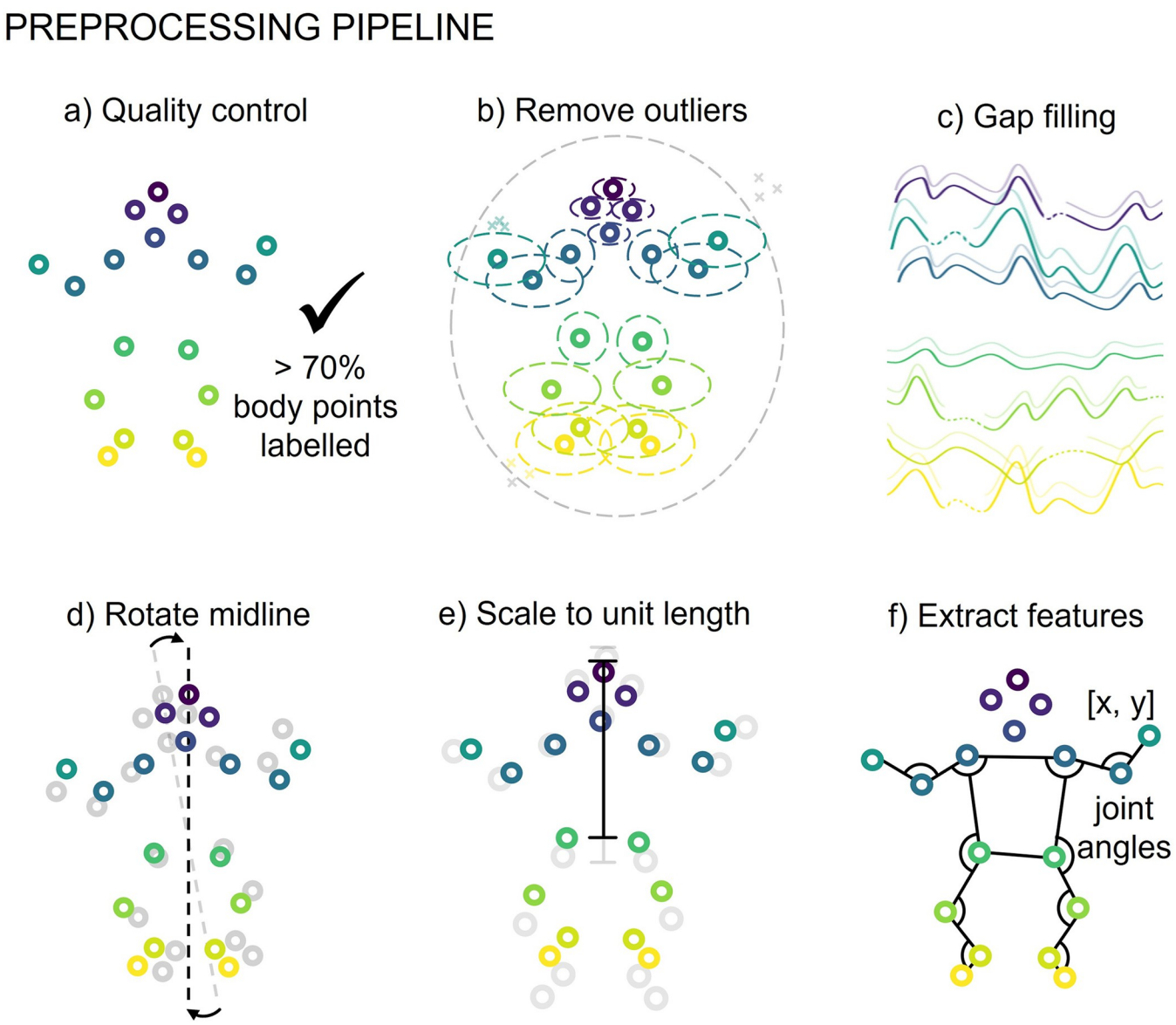


**Fig 7. Preprocessing piepline. A.** Quality control, videos with less than 70% of body points labelled on average were excluded from further analysis. **B.** Outlier removal, outlier body points were removed (denoted by x) when outside of ellipical envelope for the whole body or each body point indiviudally. **C.** Gap filling using linear interpolation for gaps less than 5 frames, or a multivariate imputer for gaps larger than 5 frames. **D.** Body points were rotated on a frame by frame basis ensuring midline of body is aligned to the vertical. **E.** Body point position was scaled to unit length based on infant size, unit length distance crown to mid hip. **F.** For each frame x- and y-coordiates and additional features consisting of joint angles from the left and right shoulder, elbow, hip, knee and ankle were extracted.

https://doi.org/10.1371/journal.pdig.0000432.g007

resulted in the exclusion of 21 videos from further analysis, resulting in a final dataset of 484 videos from 327 infants.

## Outlier removal

Mislabelling of body points may occur with automated labelling therefore, we set-up a process to remove these outliers on a frame-by-frame basis. This was done in a two-step process. First, these body points were removed using an ellipse envelope centred in the centre of the torso (mid-point of hips and shoulders). The ellipse was scaled relative to infant size, with unit

length set to distance between the infant's crown and hip midpoint. The ellipse was scaled to 3 times unit length in the proximal to distal direction and two times unit length in the medial to lateral direction. Body points lying outside of the ellipse were removed. Following this, a similar process was applied to each body point using an ellipse envelope centred at the body point's framewise median position. Each ellipse was again scaled by unit length with the proximal-distal and medial-lateral scaling set based on observed body point variance from the complete dataset. Body point labels lying outside of their respective ellipses were removed.

## Gap filling

Where gaps in body point data existed due to missing, removed, or occluded body point labels, linear interpolation was used for gaps of five frames or less. For gaps greater than five frames we used an iterative multivariate imputation [44], implemented in scikit-learn (v1.3.0) [45].

## Adjusting for camera movement

As videos were recorded on hand-held devices, camera movement relative to the infant was apparent during the three-minute video. To account for angular rotations, all points were rotated on a frame-by-frame basis so the mid-line of the body (mid-shoulder to mid-hip), was aligned to the vertical in each frame. In addition, body point position in each frame was normalised to infant unit length, measured as distance from crown to mid-hip.

## Framerate normalisation

All pre-processed movements data were normalised to the same length. Due to variation in video frame rate, the number of frames in each 3-minute video varied. To account for this, all videos were interpolated to 4500 timepoints in length or a framerate of 25 frames per second using cubic 1D interpolation as needed.

## Feature extraction

For each frame, we extracted each body point's $x$, $y$ position in addition to 10 joint angles (left and right shoulders, elbows, hips, knees and ankles; in radians, resulting in $p = 46$ features per frame ($keypoints \times \{x, y\} + joint\ angles$).

## Prediction of GMs from movement data

As abnormal GMs can occur at any point during each video, may last for different lengths of time, and occur with different frequencies, we aimed to identify short periods of time where abnormal movements were present in each video and use a sliding window approach to generate subject-level predictions (**Fig 2**). We trained a convolutional neural network to predict GMs classification. During model training, each subject's pre-processed timeseries data was split into short clips of $t = 128$ frames in length (approximately 5 sec.) with $stride = 8$. In each training epoch, we randomly sampled $s = 1$ clip per video, selecting more than one clip per video per epoch did not offer an improvement in model performance and cost more memory and computation (**Figure E in S1 Appendix**). Subject-level predictions were calculated by integrating over all clips for a given video.

## Model architecture

The model architecture is shown in **Fig 2C**. Each subject's data is represented as a tensor $S \in \mathbb{R}^{s \times t \times p}$ where $s$ is the number of sampled clips per video, $t$ is the number of frames per clip and $p$ is the number of features per frame. We employ three 1D convolutions applied along the

temporal dimension (*filters* = 64; *kernel size* = 3) with causal padding and *ReLU* activations. After each convolutional layer, we applied batch normalisation (**Fig 2C**). Each convolution was followed by max pooling along the temporal dimension with *window size* = 4 and *stride* = 4. After the final convolution, features of each clip were concatenated across the remaining time-steps to form feature matrix $M \in \mathbb{R}^{s \times 128}$ (**Fig 2C**). Clip features are then passed through a single fully-connected layer (*units* = 64, *ReLU*). We applied dropout with a rate of 0.5 before and after the connected layer. The convolutional neural network was implemented using Keras with a TensorFlow backend [46].

To identify features that discriminate subjects with or without abnormal movements, we passed each clip through a sigmoid attention module [47] (**Fig 2C**). In this context, clips with feature vectors that discriminate between classes are given a larger weight. A clip level context vector, $u$, is assigned to measure the importance of each clip to the final model output. First, each clip, $m_c \in \mathbb{R}^{1 \times 64}$, is passed though a single fully connected layer with weights and bias, $W$ and $b$, and a *tanh* activation to create clip level representation, $u_c$:

$$u_c = \tanh(Wm_c + b)$$

The similarity between each clip's representation and that of a context vector, $u$ is calculated and scaled:

$$\alpha_c = \frac{1}{1 + \exp(-u_c^T u)}$$

Where $\alpha_c \in [0,1]$ and represents the importance of each clip to the final model output. A final representation is calculated though a weighted average of clip features:

$$v = \frac{1}{n} \sum_{c=1}^{n} a_c m_c$$

Where $v$ is a feature vector representing the sampled clips from each video. The context vector, $u$, and the layer weights and biases are randomly initialised and jointly learned with other model parameters during training. The context vector, $u$, can be considered a 'signature' that identifies a discriminative movement within a clip. The resulting weighted outputs form a final feature vector, $v$. We apply a final dropout (0.5) to this vector and pass to a fully connected layer with one unit and *sigmoid* activation to predict the class label of each subject.

We used binary cross entropy (BCE) as the loss function with Stochastic Gradient Descent as the optimiser (Nesterov momentum = 0.9) [48]. As not all randomly sampled clips may contain abnormal movement patterns during each training epoch, we employed label smoothing of 0.1 to account for uncertainty in the assigned sample labels of each batch [49]. The learning rate was set to 0.005, batch size was set to 8 and we added $l_2$-regularisation of 0.005 to all weight kernels. We trained for a maximum of 10000 epochs, evaluating loss in the validation set and stopping training once validation loss had stopped improving for 100 epochs, retaining the model with minimum loss for testing.

## Data augmentation

Data augmentation is a common processing step in various image recognition and classification tasks and provides additional protection against overfitting in small sample settings [50,51]. We employed data augmentation methods for timeseries data including random magnitude scaling and time warping [50] (**Fig 2B**). We used cubic splines to generate a series of random, smooth sinusoidal curves (knots = 3–15; mean value = 1.0; sigma = 1.0). During

training: i) the timeseries in each clip were multiplied with a randomly generated curve to smoothly scale magnitude across the clip's length and ii) time warping was applied by smoothly distorting the time interval between points based on another randomly generated curve, shifting the temporal position of adjacent points closer or further apart [50] (**Fig 2B**).

## Model calibration and class imbalance

To account for the difference in class frequencies between normal and abnormal GMs (normal = 408 videos; abnormal/absent = 76 videos), we oversampled the minority class by a factor of 5 during training. For each video in the training sample with an abnormal GMs classification we sampled 5 sets of clips during each training epoch, resulting in approximately equal number of training samples from each group.

While resampling methods can improve model performance in imbalanced datasets, they can result in miscalibrated models due to the difference in class frequencies between the original sample population and the oversampled training set [52,53]. We employed Platt scaling [54] as a post-training method to calibrate model predictions. Model calibration was performed by fitting a logistic regression over model predictions in the validation dataset, the parameters of which are used to transform model outputs to calibrated probabilities at inference.

## Metadata

Age at video acquisition and birth cohort (extremely preterm or term-born) are both potential confounders that can affect GMs classification [13]. To incorporate metadata into the model, we applied an additional 1D convolution (*filters* = 4, *kernel size* = 1) to a feature vector of age at video acquisition and categorical group membership (preterm or term-born). The outputs were concatenated with the video features prior to the final layer for classification (**Fig 2C**).

## Model evaluation

At inference, each test subject's timeseries data were split into 547 overlapping clips ($t = 128$, *stride* = 8) which were passed with associated metadata through the trained and calibrated model to generate the final model output from the attention-weighted sum of all clips.

To evaluate model performance, we performed cross-validation by splitting the data into three subsets: train (70%), validate (15%) and test (15%), ensuring that the proportion of infant videos with abnormal GMs were similar across subsets. For infants with more than one video, both videos were included in the same subset. Model performance was evaluated in the test set using the area under the receiver operating curve (AUC), balanced accuracy (BA), specificity, sensitivity and positive and negative predictive values (PPV; NPV). Cross-validation was repeated 25 times, each with random splits of the dataset. Performance metrics in the test set are reported as average values across the 25 cross-validation repeats. We explore the impact of different parameter choices on model performance in the Supplemental Material (**Figure E, Figure G in S1 Appendix**). To examine important model features, we calculated model saliency for each test output using vanilla gradient maps [48].

## Baseline model

We compared model performance to alternative models based on logistic regression. For each video, we extracted a set of dynamical features previously shown to perform well in timeseries classification tasks [36], resulting in $p = 24$ features per timeseries. We concatenated timeseries features for each body point coordinate and joint angle along with associated meta data (age

and birth cohort) into a single feature vector of length = 1106. Using this data, we trained an $l_2$-regularised logistic regression model to predict GMs classification. As with the convolutional model, we performed 25 cross-validation repeats, splitting the data into 85% training and 15% testing sets. Regularisation strength was set using a grid search ($10^{-3}$–$10^{3}$) in a nested 5-fold cross-validation of the training data. To enable additional flexibility in the model, we also implemented a nonlinear kernelised logistic regression using Nystroem kernel approximation [55]. The baseline models were implemented in *scikit-learn* [45] (1.0.2), timeseries features were extracted using *pycatch22* (0.4.2) [36].

### GMs prediction and development at 2 years

Participants were followed up at 2-years' corrected age and their development assessed using the Bayley Scales of Infant and Toddler Development-3$^{rd}$ edition (Bayley-III) for motor, cognitive and language domains. Bayley-III scores were available for 252/327 infants for motor and cognitive domains and 232/327 infants for the language domain [13]. For infants with two videos, the video from the later time point was retained for comparison with Bayley-III domain scores. Each video was assigned a single GMs prediction label based on the GMs prediction label most frequently assigned during the 25-fold cross validation. This was done for each variant of model metadata inputs: movement data only (none), birth cohort, age at acquisition and combined birth and age (both). For each model variant we compared 2-year outcomes between GMs-prediction groups using an independent two sampled t-test (two-sided).

## Supporting information

**S1 Appendix. Supplementary methods and results related to article, Automated identification of abnormal infant movements from smart phone videos.**
(PDF)

## Acknowledgments

We would like to acknowledge the parents/families of infants who participated in the study. We would also like to acknowledge the extended Victorian Infant Collaborative Study team for their contribution in collecting infant and 2-year follow up data.

## Author Contributions

**Conceptualization:** E. Passmore, A. L. Kwong, J. L. Y. Cheong, A. J. Spittle, G. Ball.

**Data curation:** E. Passmore, A. L. Kwong, S. Greenstein, J. E. Olsen, A. L. Eeles, A. J. Spittle.

**Formal analysis:** E. Passmore, A. L. Kwong, S. Greenstein, A. J. Spittle, G. Ball.

**Funding acquisition:** J. L. Y. Cheong, A. J. Spittle, G. Ball.

**Investigation:** E. Passmore, A. L. Kwong, J. E. Olsen.

**Methodology:** E. Passmore, G. Ball.

**Resources:** A. L. Eeles.

**Software:** E. Passmore, S. Greenstein, G. Ball.

**Supervision:** G. Ball.

**Validation:** E. Passmore.

**Visualization:** E. Passmore, G. Ball.

**Writing – original draft:** E. Passmore, G. Ball.

**Writing – review & editing:** E. Passmore, A. L. Kwong, J. E. Olsen, A. L. Eeles, J. L. Y. Cheong, A. J. Spittle.

## References

1. Bax M, Goldstein M, Rosenbaun P, Leviton A, Paneth N, Dan B, et al. Proposed definition and classification of cerebral palsy, April 2005. Dev Med Child Neurol. 2005; 47: 571–576. https://doi.org/10.1017/s001216220500112x PMID: 16108461
2. Oskoui M, Coutinho F, Dykeman J, Jetté N, Pringsheim T. An update on the prevalence of cerebral palsy: A systematic review and meta-analysis. Dev Med Child Neurol. 2013; 55: 509–519. https://doi.org/10.1111/dmcn.12080 PMID: 23346889
3. McIntyre S, Badawi N, Balde I, Goldsmith S, Karlsson P, Novak I, et al. Australian Cerebral Palsy Register Report. 2018; 1–92.
4. Himpens E, Van Den Broeck C, Oostra A, Calders P, Vanhaesebrouck P. Prevalence, type, distribution, and severity of cerebral palsy in relation to gestational age: A meta-analytic review. Dev Med Child Neurol. 2008; 50: 334–340. https://doi.org/10.1111/j.1469-8749.2008.02047.x PMID: 18355333
5. Spittle A, Orton J, Anderson PJ, Boyd R, Doyle LW. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. Cochrane Database of Systematic Reviews. 2015;2015. https://doi.org/10.1002/14651858.CD005495.pub4 PMID: 26597166
6. Herskind A, Greisen G, Nielsen JB. Early identification and intervention in cerebral palsy. Dev Med Child Neurol. 2015; 57: 29–36. https://doi.org/10.1111/dmcn.12531 PMID: 25041565
7. Novak I, Morgan C, Adde L, Blackman J, Boyd RN, Brunstrom-Hernandez J, et al. Early, accurate diagnosis and early intervention in cerebral palsy: Advances in diagnosis and treatment. JAMA Pediatr. 2017; 171: 897–907. https://doi.org/10.1001/jamapediatrics.2017.1689 PMID: 28715518
8. Spittle AJ, Spencer-Smith MM, Cheong JLY, Eeles AL, Lee KJ, Anderson PJ, et al. General movements in very preterm children and neurodevelopment at 2 and 4 years. Pediatrics. 2013;132. https://doi.org/10.1542/peds.2013-0177 PMID: 23878041
9. Kwong AKL, Fitzgerald TL, Doyle LW, Cheong JLY, Spittle AJ. Predictive validity of spontaneous early infant movement for later cerebral palsy: a systematic review. Dev Med Child Neurol. 2018; 60: 480–489. https://doi.org/10.1111/dmcn.13697 PMID: 29468662
10. Einspieler C, Prechtl HFR. Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system. Ment Retard Dev Disabil Res Rev. 2005; 11: 61–67. https://doi.org/10.1002/mrdd.20051 PMID: 15856440
11. Christa Einspieler, Prechtl HFR. Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. Mac Keith Press; 2008.
12. Spittle AJ, Olsen J, Kwong A, Doyle LW, Marschik PB, Einspieler C, et al. The Baby Moves prospective cohort study protocol: Using a smartphone application with the General Movements Assessment to predict neurodevelopmental outcomes at age 2 years for extremely preterm or extremely low birthweight infants. BMJ Open. 2016;6. https://doi.org/10.1136/bmjopen-2016-013446 PMID: 27697883
13. Kwong AKL, Doyle LW, Olsen JE, Eeles AL, Zannino D, Mainzer RM, et al. Parent-recorded videos of infant spontaneous movement: Comparisons at 3–4 months and relationships with 2-year developmental outcomes in extremely preterm, extremely low birthweight and term-born infants. Paediatr Perinat Epidemiol. 2022; 36: 673–682. https://doi.org/10.1111/ppe.12867 PMID: 35172019
14. Byrne R, Noritz G, Maitre NL. Implementation of Early Diagnosis and Intervention Guidelines for Cerebral Palsy in a High-Risk Infant Follow-Up Clinic. Pediatr Neurol. 2017; 76: 66–71. https://doi.org/10.1016/j.pediatrneurol.2017.08.002 PMID: 28982529
15. Svensson KA, Örtqvist M, Bos AF, Eliasson AC, Sundelin HE. Usability and inter-rater reliability of the NeuroMotion app: A tool in General Movements Assessments. European Journal of Paediatric Neurology. 2021; 33: 29–35. https://doi.org/10.1016/j.ejpn.2021.05.006 PMID: 34052727
16. Adde L, Brown A, Van Den Broeck C, Decoen K, Eriksen BH, Fjørtoft T, et al. In-Motion-App for remote General Movement Assessment: A multi-site observational study. BMJ Open. 2021; 11: 1–9. https://doi.org/10.1136/bmjopen-2020-042147 PMID: 33664072
17. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Trans Pattern Anal Mach Intell. 2021; 43: 172–186. https://doi.org/10.1109/TPAMI.2019.2929257 PMID: 31331883

18. Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler P, et al. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 4929–4937. https://doi.org/10.1109/CVPR.2016.533

19. Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2014; 1653–1660. https://doi.org/10.1109/CVPR.2014.214

20. Fang H-S, Xie S, Tai Y-W, Lu C. RMPE: Regional Multi-person Pose Estimation. 2017 IEEE International Conference on Computer Vision (ICCV). 2017; 2353–2362. https://doi.org/10.1109/ICCV.2017.256

21. Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, et al. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018; 5167–5176. https://doi.org/10.1109/CVPR.2018.00542

22. Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2014; 3686–3693. https://doi.org/10.1109/CVPR.2014.471

23. Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2016; 34–50. https://doi.org/10.1007/978-3-319-46466-4_3

24. Chambers C, Seethapathi N, Saluja R, Loeb H, Pierce SR, Bogen DK, et al. Computer Vision to Automatically Assess Infant Neuromotor Risk. IEEE Trans Neural Syst Rehabil Eng. 2020; 28: 2431–2442. https://doi.org/10.1109/TNSRE.2020.3029121 PMID: 33021933

25. Groos D, Ramampiaro H, Ihlen EA. EfficientPose: Scalable single-person pose estimation. Applied Intelligence. 2021; 51: 2518–2533. https://doi.org/10.1007/s10489-020-01918-7

26. Sciortino G., Farinella G.M., Battiato S, Leo M, Distante C. Image Analysis and Processing—ICIAP 2017. Battiato S, Gallo G, Schettini R, Stanco F, editors. Cham: Springer International Publishing; 2017. https://doi.org/10.1007/978-3-319-68548-9

27. Groos D, Adde L, Støen R, Ramampiaro H, Ihlen EAF. Towards human performance on automatic motion tracking of infant spontaneous movements. 2020; 1–10. Available from: http://arxiv.org/abs/2010.05949.

28. Nguyen-Thai B, Le V, Morgan C, Badawi N, Tran T, Venkatesh S. A Spatio-Temporal Attention-Based Model for Infant Movement Assessment From Videos. IEEE J Biomed Health Inform. 2021; 25: 3911–3920. https://doi.org/10.1109/JBHI.2021.3077957 PMID: 33956636

29. Groos D, Adde L, Aubert S, Boswell L, de Regnier R-A, Fjørtoft T, et al. Development and Validation of a Deep Learning Method to Pr inedict Cerebral Palsy From Spontaneous Movements in Infants at High Risk. JAMA Netw Open. 2022; 5: e2221325. https://doi.org/10.1001/jamanetworkopen.2022.21325 PMID: 35816301

30. Reich S, Zhang D, Kulvicius T, Bölte S, Nielsen-Saines K, Pokorny FB, et al. Novel AI driven approach to classify infant motor functions. Sci Rep. 2021; 11: 1–13. https://doi.org/10.1038/s41598-021-89347-5 PMID: 33972661

31. Ihlen EAF, Støen R, Boswell L, de Regnier R-A, Fjørtoft T, Gaebler-Spira D, et al. Machine Learning of Infant Spontaneous Movements for the Early Prediction of Cerebral Palsy: A Multi-Site Cohort Study. J Clin Med. 2020; 9: 5. https://doi.org/10.3390/jcm9010005 PMID: 31861380

32. Silva N, Zhang D, Kulvicius T, Gail A, Barreiros C, Lindstaedt S, et al. The future of General Movement Assessment: The role of computer vision and machine learning–A scoping review. Res Dev Disabil. 2021;110. https://doi.org/10.1016/j.ridd.2021.103854 PMID: 33571849

33. Redd CB, Karunanithi M, Boyd RN, Barber LA. Technology-assisted quantification of movement to predict infants at high risk of motor disability: A systematic review. Res Dev Disabil. 2021;118. https://doi.org/10.1016/j.ridd.2021.104071 PMID: 34507051

34. Irshad MT, Nisar MA, Gouverneur P, Rapp M, Grzegorzek M. AI approaches towards prechtl's assessment of general movements: A systematic literature review. Sensors (Switzerland). MDPI AG; 2020. pp. 1–32. https://doi.org/10.3390/s20185321 PMID: 32957598

35. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018; 21: 1281–1289. https://doi.org/10.1038/s41593-018-0209-y PMID: 30127430

36. Lubba CH, Sethi SS, Knaute P, Schultz SR, Fulcher BD, Jones NS. catch22: CAnonical Time-series CHaracteristics: Selected through highly comparative time-series analysis. Data Min Knowl Discov. 2019; 33: 1821–1852. https://doi.org/10.1007/s10618-019-00647-x

37. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2nd International Conference on Learning Representations, ICLR 2014—Workshop Track Proceedings. 2014; 1–8.

38. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Adv Neural Inf Process Syst. 2018;2018-Decem: 9505–9515.

39. Spittle AJ, Spencer-Smith MM, Eeles AL, Lee KJ, Lorefice LE, Anderson PJ, et al. Does the Bayley-III Motor Scale at 2 years predict motor outcome at 4 years in very preterm children? Dev Med Child Neurol. 2013; 55: 448–452. https://doi.org/10.1111/dmcn.12049 PMID: 23216518

40. Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc. 2019; 14: 2152–2176. https://doi.org/10.1038/s41596-019-0176-0 PMID: 31227823

41. Lachance J, Thong W, Nagpal S, Xiang A. The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R 2 HCAI) A Case Study in Fairness Evaluation: Current Limitations and Challenges for Human Pose Estimation. Available from: https://scholar.google.com.

42. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009; 42: 377–381. https://doi.org/10.1016/j.jbi.2008.08.010 PMID: 18929686

43. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019; 95: 103208. https://doi.org/10.1016/j.jbi.2019.103208 PMID: 31078660

44. van Buuren S, Oudshoorn CGM. MICE: multivariate imputation by chained equations. R package version. 2007; 1: 2007.

45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–2830.

46. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available from: https://www.tensorflow.org/

47. Yang Y, Sun J, Li H, Xu Z. Deep ADMM-Net for compressive sensing MRI. Adv Neural Inf Process Syst. 2016; 10–18.

48. Sutskever I, Martens J, Dahl G, Hinton G. Momentum, Nesterov accelerate, On the importance of initialization and momentum in deep learning. J Mach Learn Res. 2013; 28: 1139–1147.

49. Zhang C Bin, Jiang PT, Hou Q, Wei Y, Han Q, Li Z, et al. Delving deep into label smoothing. IEEE Transactions on Image Processing. 2021; 30: 5984–5996. https://doi.org/10.1109/TIP.2021.3089942 PMID: 34166191

50. Um TT, Pfister FMJ, Pichler D, Endo S, Lang M, Hirche S, et al. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. Proceedings of the 19th ACM International Conference on Multimodal Interaction. 2017; 216–220. https://doi.org/10.1145/3136755.3136817

51. Salamon J, Bello JP. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Process Lett. 2017; 24: 279–283. https://doi.org/10.1109/LSP.2017.2657381

52. He H, Garcia EA. Learning from imbalanced data. Studies in Computational Intelligence. 2019; 807: 81–110. https://doi.org/10.1007/978-3-030-04663-7_4

53. Aggarwal U, Popescu A, Belouadah E, Hudelot C. A comparative study of calibration methods for imbalanced class incremental learning. Multimed Tools Appl. 2022; 81: 19237–19256. https://doi.org/10.1007/s11042-020-10485-5

54. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999; 10: 61–74.

55. Seeger M C. Williams. Using the Nyström method to speed up kernel machines. Advances in Neural Information Processing Systems. 2001. pp. 682–688.