Towards Scalable Newborn Screening: Automated General Movement Assessment in Uncontrolled Settings

Daphné Chopard* DCHOPARD@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland
Department of Intensive Care and Neonatology and Children's Research Center, University of Zurich, University Children's Hospital Zürich, Zurich, Switzerland

Sonia Laguna* SLAGUNA@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland

Kieran Chin-Cheong* KCHINCHEONG@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland

Annika Dietz
Department of Neonatology, University Children's Hospital Regensburg, Hospital St. Hedwig of the Order of St. John, University of Regensburg, Regensburg, Germany

Anna Badura
Department of Neonatology, University Children's Hospital Regensburg, Hospital St. Hedwig of the Order of St. John, University of Regensburg, Regensburg, Germany

Sven Wellmann
Department of Neonatology, University Children's Hospital Regensburg, Hospital St. Hedwig of the Order of St. John, University of Regensburg, Regensburg, Germany

Julia E Vogt JULIA.VOGT@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland

Abstract

General movements (GMs) are spontaneous, coordinated body movements in infants that offer valuable insights into the developing nervous system. Assessed through the Prechtl GM Assessment (GMA), GMs are reliable predictors for neurodevelopmental disorders. However, GMA requires specifically trained clinicians, who are limited in number. To scale up newborn screening, there is a need for an algorithm that can automatically classify GMs from infant video recordings. This data poses challenges, including variability in recording length, device type, and setting, with each video coarsely annotated for overall movement quality. In this work, we introduce a tool for extracting features from these recordings and explore various machine learning techniques for automated GM classification.

1. Introduction

General movements (GMs) refer to spontaneous movements of the entire body observable in infants from early fetal life until approximately six months post-term (Einspieler and Prechtl, 2005). These movements include coordinates sequences of the arms, legs, neck, and trunk, with variations in intensity, force, and speed. GMs are inherently complex, variable, and fluid, and their quality can provide critical insights into the integrity of the

* Equal contribution

developing nervous system. Assessing GM quality, known as the Prechtl General Movement Assessment (GMA), is a sensitive and reliable diagnostic tool for predicting the likelihood of cerebral palsy and other neurodevelopmental disorders (Einspieler and Prechtl, 2005). The types of movements predictive of later neurodevelopmental disorders vary significantly with age, and age-based distinctions are essential when studying newborns. From early fetal life until approximately two months post-term, GMs exhibit consistent writhing movements. Between 6 to 9 weeks post-term, these writhing patterns diminish, and fidgety movements emerge, persisting until around the middle of the first year of life, at which point intentional movements become predominant. Alterations in these movement patterns can indicate later disorders. While GMA is typically conducted in hospitals by specialized clinicians, there are too few trained experts to screen all newborns. This motivates the need to develop a tool for automatic and reliable classification of movements.

Initial attempts at addressing this challenge focused on computer-based approaches using feature extraction techniques (Baccinelli et al., 2020; Raghuram et al., 2022). However, recent advances in deep learning and computer vision have spurred the development of more sophisticated methods (Silva et al., 2021; Irshad et al., 2020). For instance, Reich et al. (2021) proposed to use a pose estimator, OpenPose, to extract skeletons and used these as input to a shallow multilayer neural network to discriminate between movements. Other deep learning solutions, e.g. by Schmidt et al. (2019), work in controlled conditions. Despite their promise, these approaches often rely on controlled environments with fixed sensors and constant recording length, limiting their scalability and applicability in real-world clinical settings.

In this work, we present a preliminary study introducing a method to label key anatomical points, automatically process videos, and classify GMs based on the tracked anatomical points across videos, using a dataset from the Barmherzige Brüder Regensburg Hospital. The challenges posed by the variability in recording lengths and devices, the diversity of video scenarios (including both hospital and home environments), and the nature of the movement quality annotations—one per recording—are significant for reliable GM classification. Addressing these challenges is crucial to the generalization of an automatic GMA that can be effectively deployed in varied settings. The goal is to enable comprehensive newborn screening and follow-up studies by experts, ultimately improving the early detection and treatment of neurodevelopmental disorders.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work provides insight into the feasibility of building screening tools for GMA from videos captured in an uncontrolled environment. Unlike prior studies conducted in highly controlled environments, our dataset reflects clinical reality: varied recording devices, environments, and video quality. We show that even with coarse, single-label annotations per video—a common limitation in clinical data—robust classification of infant movement quality is achievable. Our findings emphasize the value of carefully designed preprocessing pipelines and simple, computationally efficient models, which perform competitively despite limited data. Additionally, the performance gap between early and late infancy groups suggests that age-specific modelling strategies may be necessary, highlighting the importance of aligning machine learning approaches with developmental context.

2. Dataset

This study includes 76 infant video recordings collected from the Barmherzige Brüder Regensburg Hospital. The dataset is highly heterogeneous, including videos recorded both at home and in hospital settings using various recording devices, resulting in differing resolutions and frame rates. There is also significant variability in camera angles and orientations, distance from subject to camera, infants' clothing, and background types, among other factors.

The data is divided into two age-based groups according to the GM phase. The *early General Movement* group includes 39 preterm infants recorded after birth at postmenstrual ages between 32 and 36 weeks (mean: 33 weeks), while the *late General Movement* group comprises 37 infants with postmenstrual ages ranging from 49 to 59 weeks (mean: 53 weeks). Each sample is associated with a binary label reflecting the infant's movement quality. Following Einspieler and Prechtl (2005), in the *early GM* group, which encompasses preterm and writhing GMs, labels distinguish between normal and poor movement repertoire. In contrast, the *late GM* group labels indicate the presence or absence of fidgety movements, with the presence of fidgety movements considered normal.

To minimize bias and improve reliability, two trained physicians independently annotated the videos. In cases of disagreement, the recordings were reviewed jointly until a consensus label was reached. In the dataset, 24 infants in the early GM group (65%) exhibit poor repertoire, while 10 infants in the late GM group (27%) lack fidgety movements. These proportions are consistent with previous clinical studies in high-risk populations Sæther et al. (2016); De Bock et al. (2017); Alonzo et al. (2022), although notably higher than the prevalence of about 3% reported in cohorts of general population Wu et al. (2020).

Table 1 shows several demographic characteristics, as well as details about the videos themselves for both the early and late GM groups. Unfortunately, some details are not available to us, such as weight at video recording time for early GM group infants. Figure 1 provides histograms for the distributions of both video resolution and FPS for the early and late GM groups. Note that not all video segments are equally informative: while GMA usually requires 2 to 5 minutes of video, 15 to 30 seconds of relevant movement may be sufficient Kapil et al. (2024).

Table 1: Demographic and video recording characteristics of infants in both GM groups.

Characteristic	**Early GM Group**	**Late GM Group**
Age at video recording [days] (mean ± SD)	30.9 ± 19.9	170.4 ± 24.8
Post-menstrual age at video recording [days] (mean ± SD)	231.5 ± 6.5	N/A
Corrected Age at video recording [days] (mean ± SD)	N/A	92.5 ± 17.4
Weight at video recording [g] (mean ± SD)	N/A	4148.2 ± 830.7
Video Length [s] (mean ± SD)	409 ± 144	224 ± 93
FPS [min, max]	[30, 120]	[30, 120]
Extremely Preterm	16	15
Very Preterm	20	19
Moderate to Late Preterm	3	3
Not Preterm	0	0

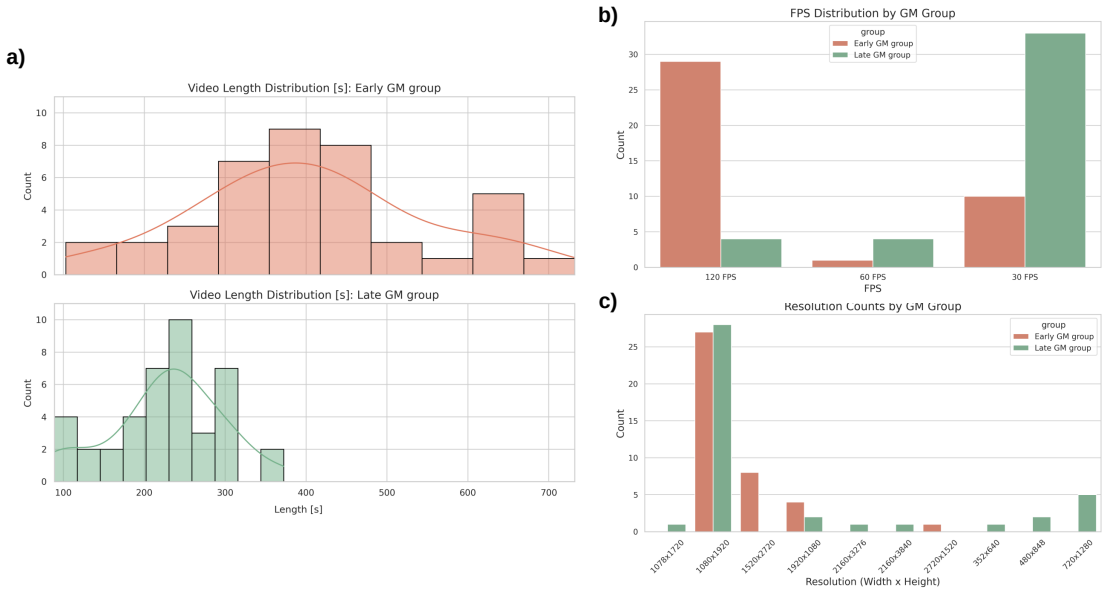

Figure 1: Histograms showing the distributions of (a) video length, (b) frame rate per second (FPS), and (c) video resolution for each GM group. The videos are highly heterogeneous, with substantial variation in recording length, frame rate, and resolution. FPS values are rounded to the closest one.

3. Methods

This work involves extracting features from specific anatomical landmarks of interest—so-called *keypoints (KP)*—in video frames (Section 3.3) to classify movement quality. We explore two means of acquiring the desired keypoints: 1) ***Label & Track***, manually labelling specific KPs in a video frame (Section 3.1), and tracking their positions across the video (Section 3.1), and 2) ***AggPose***, automatically labeling the KPs in each frame, using the existing ML labeller AggPose (Cao et al., 2022). The overall pipeline is illustrated in Figure 2. The code[1] is publicly available.

3.1. Keypoint Retrieval Tools

In this section, we provide details about the two approaches for extracting keypoints from videos. In the first approach, *Label & Track*, keypoints are manually labelled in a single frame and then tracked throughout the rest of the video. In the second approach, *Automatic Label*, keypoints are automatically labelled in each video frame independently.

3.1.1. Label & Track

Keypoint Labelling We introduce a keypoint labelling tool based on the work from Doersch et al. (2023) to mark sets of key coordinates in the videos, which, along with point

1. https://github.com/mdslabeth/GMA

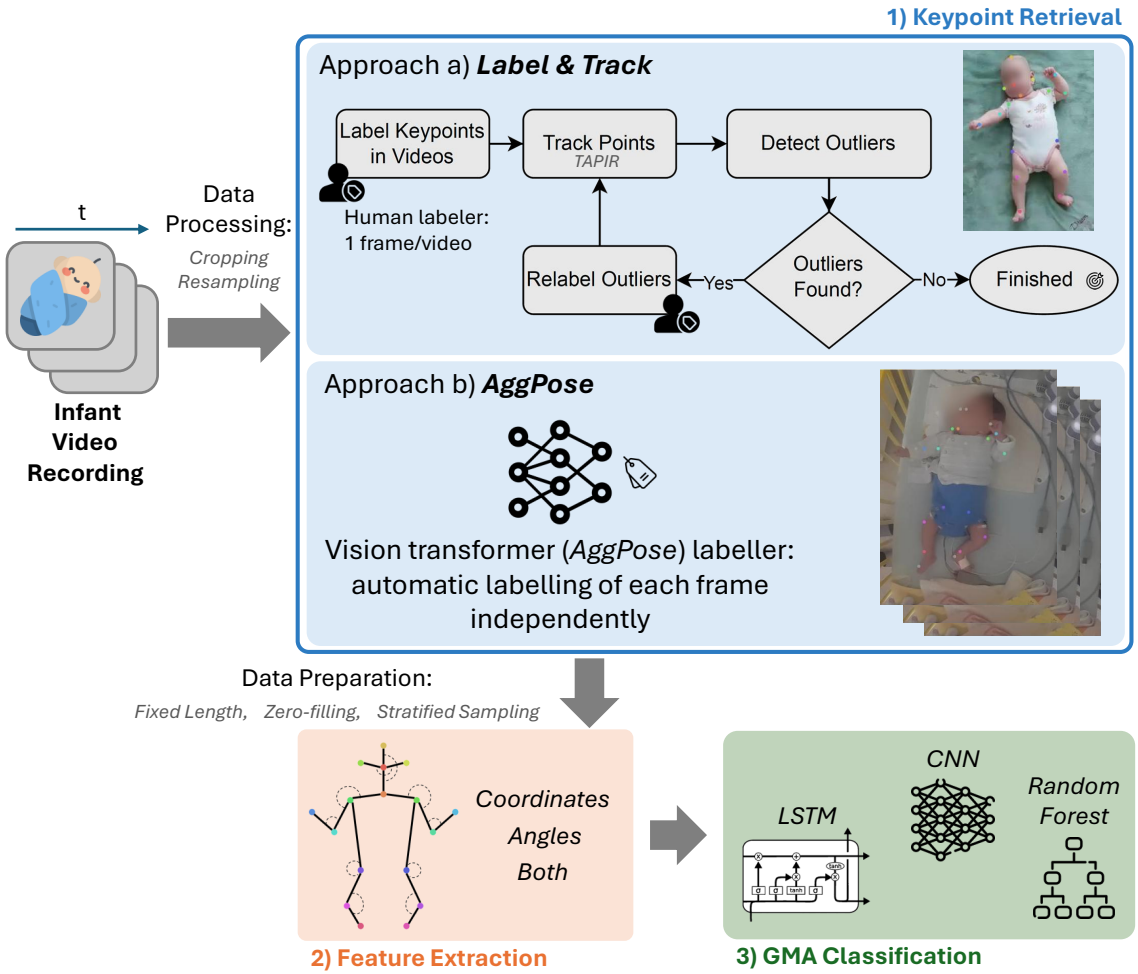

Figure 2: Overview of the GMA classification pipeline. Video recordings undergo preprocessing steps (Section 3.2). Keypoints are then extracted using one of two approaches described in Section 3.1: (a) **Label & Track**, or (b) **AggPose**. Features are extracted (Section 3.3) from the keypoints and used to train GMA classification models (Section 3.4) including CNN, LSTM, and Random Forest for prediction.

tracking, forms the backbone of the preprocessing pipeline. The tool allows labelling of extremities such as wrists and ankles—referred to as *extreme keypoints*, as well as additional joints such as elbows, hips, and knees—collectively referred to as *all keypoints*, totaling 17. In this study, three independent non-expert labellers manually labelled all keypoints for subsequent tracking. Full lists of relevant keypoints can be found in Figure 4.

Point Tracking For tracking labelled keypoints across video frames, we utilize the TAPIR point tracking algorithm proposed by Doersch et al. (2023). TAPIR, is designed to track arbitrary physical points on surfaces over video sequences by combining per-frame initial-

ization with temporal refinement. It begins by computing a cost volume (namely a 4D map with dimensions corresponding to time, height, width, and channels) that captures the similarity between the feature representation of a query point and all possible spatial features in each frame. This cost volume is used to generate an initial estimate of the point's position, along with associated occlusion and uncertainty metrics. This is followed by an iterative refinement stage, where the local spatio-temporal features are processed using a depthwise-convolutional network to improve tracking accuracy over time. TAPIR is trained on a modified version of the Kubric MOVi-E dataset Greff et al. (2022), which contains simulated videos featuring physically realistic interactions between deformable and rigid objects. Training supervision includes Huber loss for point location regression and binary cross-entropy loss for occlusion and uncertainty prediction. TAPIR is evaluated on TAP-Vid Doersch et al. (2022), a collection of four diverse datasets (including TAP-Vid-Kinetics and TAP-Vid-DAVIS) each posing distinct tracking challenges. It achieves state-of-the-art performance and demonstrates strong robustness in occlusion-heavy and dynamic scenes.

Figure 3: Consecutive frames with unrealistic keypoint tracking. The outlier detection tool displays potential unrealistic point movements and allows manual correction.

3.1.2. AggPose: Automatic label

As a benchmark, we also evaluate a fully automated keypoint extraction method: AggPose (Cao et al., 2022), a vision transformer model specifically designed for infant pose estimation and trained on a large-scale infant pose dataset. AggPose has been shown to outperform established models such as OpenPose Cao et al. (2019) and HRNet Wang et al. (2020), making it a strong and representative baseline for this class of methods. Unlike our *Label & Track* pipeline, AggPose eliminates the need for both manual labelling and tracking by independently predicting keypoints in each frame.

Unlike traditional convolutional models, AggPose discards convolutional layers in favour of a fully transformer-based architecture, leveraging multi-scale feature aggregation through

a deep-layer fusion mechanism. This method enhances spatial information sharing across different resolution levels, improving the robustness of KP detection in challenging infant pose scenarios. We provide a comprehensive overview of this method in Appendix A.

In this study, we use AggPose to automatically label KPs in each frame of the video dataset, extracting the full set of coordinates across time. This method encompasses a total of 21 KPs, slightly larger than the prior approach. See Figure 2 Approach b) for a visualization of the KPs, compared to manual labelling, some keypoints are omitted (head top, ears, and nose), while others are added (eyes, hand and foot extremities, torso)– displayed in grey–when labelling automatically.

3.2. Data Preprocessing

Due to the heterogeneity of the data, a robust preprocessing and error correction pipeline is necessary to prepare the videos for machine learning model development. This subsection outlines the key steps in that process, with a graphical overview in Figure 2 Approach a.

Resampling and Extreme Keypoints Labelling Preprocessing begins with homogenizing all the video data by resampling to 30 frames per second. Extreme keypoints are then labelled. If all required keypoints are not visible in the first frame, labelling occurs in a subsequent frame where visibility is clearer.

Video Cropping Next, the extracted extreme keypoints are tracked and used to apply a rectangular crop per video. The crop is determined by the most extreme values of the tracked points in each video, with a 15% margin to ensure the infant consistently occupies a similar portion of the video frame in all videos. This normalization step is crucial for maintaining consistency in the input data for machine learning models.

Outlier Detection To address inaccuracies in the tracking algorithm in the *Label & Track* (See 3.1.1) approach, such as when points jump inappropriately due to occlusions or overlapping limbs, an outlier detection mechanism is applied.

The tool automatically flags any sudden and unrealistic movement of keypoints and prompts manual relabelling by the annotators, with an example shown in Figure 3. Specifically, we define such movements as coordinate changes that exceed 15 times the overall standard deviation for each keypoint. This threshold is set deliberately high to avoid flagging normal, natural movements, as well as fidgety movements that can inherently exhibit high magnitudes without indicating an error. After relabelling, the affected keypoints are retracked. This process is iterated a fixed number of rounds, two in these results, as we observed only marginal improvements in performance after three rounds in the final detection task, or until no further outliers are detected, ensuring the accuracy of the tracked coordinates. In general, this outlier-detection process can be structured to run for a set number of iterations, until a defined performance tolerance is reached, or until no further outliers are identified, ensuring the accuracy of the tracked coordinates.

Final Data Preparation Prior to classification, we homogenize the time series data. After resampling to 30 fps, each video is split into fixed-length sequences equal to the shortest recording within its respective age group (early: 616 frames; late: 674 frames; approximately 20.5 s and 22.5 s, respectively). Zero-filling is applied to any occluded or

missing keypoint coordinates. To prevent data leakage, we use stratified sampling to ensure that no infant appears in both the training and testing datasets. If a video contains multiple full-length clips, each is treated as a separate instance; incomplete final fragments are discarded.

3.3. Feature Extraction

We consider three feature sets from the tracked keypoints to perform the classification task: x- and y-coordinates of all keypoints and the angles between selected keypoints, based on Prakash et al. (2023).

Figure 4: List of *all* and *extreme* keypoints considered in this study for manual labelling, coloured by label on the right.

All Keypoints	Extreme
nose	
head bottom	
head top	x
left ear	
right ear	
left shoulder	
right shoulder	
left elbow	x
right elbow	x
left wrist	x
right wrist	x
left hip	
right hip	
left knee	x
right knee	x
left ankle	x
right ankle	x

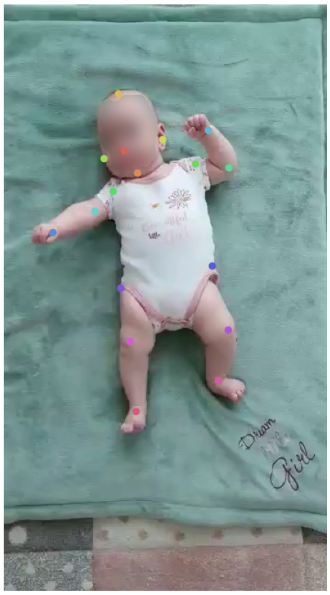

Figure 5: Example of frame with labelled keypoints.

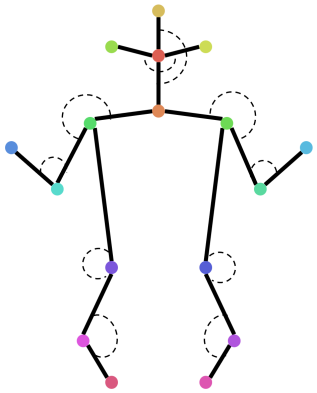

Figure 6: Illustration of the angle features. The dash lines show the angles that we compute in the set of angle features.

Keypoint Coordinates The x- and y-coordinates of all labelled keypoints are treated independently as separate input channels, and used as a simple representation of the keypoint positions across frames, as listed in Figure 4 and Figure 5.

Keypoint Angles The angles are calculated between keypoint triplets to capture movement dynamics. Given three keypoints $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ with coordinates (x_1, y_1), (x_2, y_2)

and (x_3, y_3) respectively, we compute the angle θ formed at middle point $\mathbf{p}_2$ between vectors $\mathbf{v}_1 = \mathbf{p}_1 - \mathbf{p}_2$ and $\mathbf{v}_2 = \mathbf{p}_3 - \mathbf{p}_2$, defined as

$$\mathbf{v}_1 = \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} x_3 - x_2 \\ y_3 - y_2 \end{bmatrix} \tag{1}$$

$$\theta = \arccos\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right). \tag{2}$$

To ensure numerical stability, the cosine value is clipped to the range $[-1, 1]$ before applying the arccos function. Here, $\cdot$ denotes the dot product and $\|\cdot\|$ the Euclidean norm.

Table 2 lists the selected angle sets and the anatomical keypoints involved, with intuitive descriptions. These are illustrated in Figure 6. Although the automatic labelling method predicts more keypoints, the same set of angles is used.

$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$	**Explanation**
head top	nose	head bottom	Head top to neck angle
right ear	nose	left ear	Right to left ear angle
left elbow	left shoulder	head bottom	Head to left shoulder angle
right elbow	right shoulder	head bottom	Head to right shoulder angle
left wrist	left elbow	left shoulder	Left elbow angle
right wrist	right elbow	right shoulder	Right elbow angle
left knee	left hip	left shoulder	Left hip angle
right knee	right hip	right shoulder	Right hip angle
left hip	left knee	left ankle	Left knee angle
right hip	right knee	right ankle	Right knee angle

Table 2: List of angle components used in the study with intuitive explanations.

Both The combination of both coordinates and angles, shown as "both" in the results.

3.4. Classification Models

This work explores three different classification models for GMA, using the extracted time series described in Section 3.3. 5-fold stratified cross-validation (CV) is performed across models to ensure robustness.

1D Convolutional Neural Network (1D-CNN) 1D-CNNs are a variant of the more common 2D convolutional neural network, which apply convolution operations to 1D signals, making them well-suited for processing time-series data. These networks have been shown to perform well when working with smaller labelled datasets and for specific applications (Kiranyaz et al., 2021).

Long Short-Term Memory (LSTM) LSTM networks (Hochreiter and Schmidhuber, 1997) are a type of recurrent neural network capable of exploiting the temporal aspect of our data by learning relevant context information from previously seen time points. LSTMs

are ideal for handling time series and assessing whether longer temporal memory helps in recognizing and classifying infant's general movements.

Random Forest (RF) RFs (Breiman, 2001) are powerful discriminative classifiers that perform well on a variety of datasets with minimal tuning (Biau and Scornet, 2016). It is an ensemble method that builds multiple decision trees trained on a random subset of data and combines their outputs to improve prediction accuracy and reduce overfitting. Although they do not explicitly account for the temporal structure of the data, their efficiency and accuracy make them a strong candidate for our classification tasks.

4. Experiments, Results and Discussion

4.1. Implementation details

We performed a grid search over the hyperparameter space for each of the three classifier types in Section 3.4, choosing the values that performed best in the evaluation. The final models were trained using these hyperparameters with 5-fold stratified CV, and the results from the held-out test set, with 20% of the data, were collected. On the *Label & Track* approach, the classification models used included a 1D-CNN with binary cross entropy (BCE) loss, learning rate of 0.00001, batch size 6, a fully-connected bottleneck with 150 features, and trained for 150 epochs. The LSTM used the BCE loss, learning rate 0.001, batch size 6, 3 layers, a hidden size of 64 and trained for 200 epochs. Finally, the random forest had 170 estimators. Moreover, using the *Aggpose* approach, the 1D-CNN had a learning rate of 0.0001, batch size of 4 and 150 features in the fully connected bottleck trained for 50 epochs. The LSTM also used BCE loss learning rate 0.0001, batch size 4, 2 layers, a hidden size of 64 and trained for 50 epochs. Lastly, the random forest in this setup had 45 estimators. All experiments were run using 30 independent seeds on each individual set. Our models were implemented using pytorch v2.3.1 and scikit-learn v1.3.0. The final classification was evaluated using accuracy, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision–Recall curve (AUPRC).

4.2. Results and Discussion

Table 3 shows the results of the different classification models, across the different keypoint retrieval methods as measured in AUROC, AUPRC and Accuracy. See Appendix B and C for a comprehensive overview of the results, which are also broken down by individual keypoint labeller performance. They demonstrate a variety of outcomes across age groups and feature sets, with no clear pattern indicating that one method is superior. However, the use of angle features, either isolated or combined, proves useful, as the angles only or both sets of input features generally outperform the coordinates only set, sometimes quite significantly. Additionally, the early GM group shows significantly better overall results, this indicates a large difference in the difficulty of the task between the age groups, given that we have roughly the same number of videos for each age group. Overall, the *Label & Track* approach performs comparably to the automatically labeled *AggPose* method. However, in a significant number of data groups, it demonstrates improved results, highlighting the positive impact of keypoint tracking. By incorporating temporal information from the time series to determine the position of keypoints at each time frame, this approach enhances

Table 3: Performance of various classifiers for GM classification across the different keypoint retrieval methods. The mean performance across 30 seeds and 3 keypoint labellers is reported for the *Label & Track* method, and the mean and standard deviation over 30 seeds for *AggPose* is reported. The best result per model and method is bolded.

Metric	Keypoint Retrieval	Input Features	RF	LSTM	CNN
Early GM Group					
AUROC	*Label & Track*	Coordinates	0.58	**0.65**	0.72
		Angles	**0.77**	0.62	0.75
		Both	0.70	0.65	**0.75**
	AggPose	Coordinates	0.48 ± 0.19	0.57 ± 0.19	0.46 ± 0.18
		Angles	$\mathbf{0.51} \pm 0.19$	$\mathbf{0.61} \pm 0.21$	0.49 ± 0.15
		Both	0.48 ± 0.17	0.53 ± 0.21	$\mathbf{0.51} \pm 0.19$
AUPRC	Label & Track	Coordinates	0.72	**0.74**	**0.84**
		Angles	**0.84**	0.70	0.80
		Both	0.79	0.73	0.84
	AggPose	Coordinates	0.63 ± 0.16	0.70 ± 0.15	0.65 ± 0.15
		Angles	$\mathbf{0.68} \pm 0.15$	$\mathbf{0.73} \pm 0.16$	0.65 ± 0.13
		Both	0.63 ± 0.16	0.65 ± 0.16	$\mathbf{0.68} \pm 0.15$
Accuracy	*Label & Track*	Coordinates	0.54	0.60	0.66
		Angles	**0.69**	0.57	0.67
		Both	0.60	**0.63**	**0.67**
	AggPose	Coordinates	0.47 ± 0.13	0.54 ± 0.16	0.47 ± 0.15
		Angles	$\mathbf{0.50} \pm 0.17$	$\mathbf{0.59} \pm 0.14$	0.47 ± 0.11
		Both	0.47 ± 0.15	0.53 ± 0.20	$\mathbf{0.49} \pm 0.14$
Late GM Group					
AUROC	*Label & Track*	Coordinates	0.45	0.45	**0.57**
		Angles	**0.59**	**0.54**	0.55
		Both	0.48	0.47	0.56
	AggPose	Coordinates	0.40 ± 0.26	0.54 ± 0.23	0.56 ± 0.24
		Angles	$\mathbf{0.44} \pm 0.25$	$\mathbf{0.66} \pm 0.24$	$\mathbf{0.72} \pm 0.24$
		Both	0.40 ± 0.24	0.65 ± 0.26	0.52 ± 0.29
AUPRC	*Label & Track*	Coordinates	0.31	0.36	**0.50**
		Angles	**0.45**	**0.42**	0.39
		Both	0.33	0.35	0.43
	AggPose	Coordinates	0.36 ± 0.21	0.50 ± 0.24	0.50 ± 0.25
		Angles	$\mathbf{0.36} \pm 0.19$	0.58 ± 0.27	$\mathbf{0.65} \pm 0.28$
		Both	0.33 ± 0.18	$\mathbf{0.59} \pm 0.28$	0.45 ± 0.28
Accuracy	*Label & Track*	Coordinates	0.63	0.59	0.60
		Angles	**0.67**	**0.64**	0.62
		Both	0.64	0.62	**0.62**
	AggPose	Coordinates	0.54 ± 0.17	0.60 ± 0.17	0.62 ± 0.18
		Angles	$\mathbf{0.62} \pm 0.11$	$\mathbf{0.71} \pm 0.15$	$\mathbf{0.70} \pm 0.18$
		Both	0.59 ± 0.14	0.61 ± 0.18	0.61 ± 0.19

classification performance, whereas *AggPose* labels each keypoint in isolated frames, missing potential contextual cues. Furthermore, the point tracking algorithm used by *Label & Track* is explicitly trained to handle keypoint occlusions, which occur often in these videos. Finally, the *Label & Track* approach benefits from the two rounds of manual keypoint correction for detected outliers, as described in Section 3.2. See Appendix C for more information regarding the relative performance between *Label & Track* and *AggPose*, including a breakdown of which experiment settings show statistically significant differences. One of the key takeaways from this study is that despite the relatively small and heterogeneous dataset, our approach has shown success in automatically classifying GM in newborns, with average AUROCs of up to 0.8283 in certain scenarios, opening promising lines of research for GMA "in the wild".

5. Conclusion, Limitations and Future Outlook

This study presents a preliminary but promising step toward the development of an automated tool for General Movements Assessment (GMA). We demonstrate that infant motor quality can be predicted from video recordings with an AUROC of up to 0.8283, with overall better performance in the early GM group. This highlights the potential of using automated methods to assess infant motor behaviour. Despite the challenges posed by diverse data sources and recording conditions in the data, our results show that automated GMA is feasible in real-world clinical settings. This is a significant step forward in the early detection of neurodevelopmental disorders. This study lays the groundwork for further advancements, emphasizing the need for larger datasets and more refined models and feature extraction approaches to capture the data complexities. Ultimately, the goal is to enhance early prediction of neurodevelopmental diseases, offering a valuable tool for clinicians, and ultimately improving patient care.

Limitations One of the key limitations of this study is the small dataset size, which affects both the robustness and generalizability of the results, as reflected in the relatively large standard deviations across cross-validation folds. This also restricts our ability to perform stratified error analyses, such as evaluating performance differences across age groups or recording conditions, both of which are clinically relevant due to the evolving nature of infant movements.

The demographic homogeneity of the dataset is another constraint, particularly in assessing how well the model performs across subpopulations. For example, evaluating the effect of skin tone on keypoint tracking is not feasible with the current cohort. While our *Label Track* approach relies on a tracking algorithm that uses general visual features and is not explicitly tied to skin appearance, this assumed robustness must still be empirically validated. Importantly, the modular nature of our pipeline allows for the integration of improved tracking algorithms as the field advances.

The heterogeneity of the video data (including differences in camera quality, angles, environments, and frame rates) adds complexity to the task but also increases clinical relevance by reflecting real-world variability. However, the limited dataset size again restricts our ability to systematically analyse how these factors influence model performance.

A further limitation is the coarse, video-level annotation of movement quality. In practice, not all segments of a video are equally informative. Since we segment longer recordings

into fixed-length clips and treat them independently, the use of a single video-level label introduces noise, especially when only a portion of the video reflects abnormal movements. Finer-grained, segment-level annotations would reduce this label noise and likely improve model accuracy. While such annotations would introduce new challenges (e.g., class imbalance), they would also enable more advanced approaches like Multiple Instance Learning (MIL), which could increase interpretability by focusing on which segments lead to a particular classification. Additionally, since GMA is assessed on a continuous scale, integrating that granularity into modelling could provide richer predictions provided a larger dataset.

Finally, the small dataset also limits our ability to explore more powerful models, such as Vision Transformers (such as TimeSformer), which may better capture the complex spatiotemporal structure of infant movements.

Future Outlook Future work on this dataset could focus on engineering more sophisticated features from the extracted keypoints particularly those that capture temporal dynamics relevant to infant motor behavior. Exploring more powerful model classes–such as Transformers–may also better leverage the sptial and temporal patterns present in the videos. Addionally, our current outlier detection correction pipeline, used for manually labeled keypoints, could be used to refine the keypoints extracted by the *AggPose* automatic keypoint retrieval method. Replacing the current tracking module in the *Label & Track* approach with more recent point tracking algorithms could also further improve robustness. As the dataset grows, more comprehensive analyses will become possible, including stratified error analysis across different age groups, recording conditions, and demographic subgroups. Interpretability remains a critical goal in medical applications; identifying movement patterns that drive classification decisions would be highly valuable. With more fine-grained or continuous labels, approaches such as Multiple Instance Learning could offer more transparent and clinically relevant predictions. While GMA is a standardized assessment, inter-institutional studies of annotator agreement would further validate model generalizability. Ultimately, if predictive performance can be maintained or improved, the development of a decision support tool to assist clinicians in screening more infants using GMA would be a highly impactful application of this work.

Acknowledgments

DC is funded from the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" grant #2021-911 of the ETH Domain (Swiss Federal Institutes of Technology), and SL and KC from the Swiss State Secretariat for Education, Research and Innovation (SERI), contract #MB22.00047. The authors would like to thank Ričards Marcinkevičs and Heike Leutheuser for their insights and discussions on the development of this manuscript and all the reviewers for their helpful feedback and suggestions.

References

Corrie J Alonzo, Lisa C Letzkus, Elizabeth A Connaughton, Nancy L Kelly, Joseph A Michel, and Santina A Zanelli. High prevalence of abnormal general movements in hospitalized very low birth weight infants. *American Journal of Perinatology*, 29(14):1541–1547, 2022.

Walter Baccinelli, Maria Bulgheroni, Valentina Simonetti, Francesca Fulceri, Angela Caruso, Letizia Gila, and Maria Luisa Scattoni. Movidea: A software package for automatic video analysis of movements in infants at risk for neurodevelopmental disorders. *Brain sciences*, 10(4):203, 2020.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

L Breiman. Random forests. *Machine Learning*, 2001.

Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, and Jianguo Cao. Aggpose: Deep aggregation vision transformer for infant pose estimation. *arXiv preprint arXiv:2205.05277*, 2022.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

Freia De Bock, Heike Will, Ulrike Behrenbeck, Marc N Jarczok, Mijna Hadders-Algra, and Heike Philippi. Predictive value of general movement assessment for preterm infants' development at 2 years- implementation in clinical routine in a non-academic setting. *Research in Developmental Disabilities*, 62:69–80, 2017.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023.

Christa Einspieler and Heinz FR Prechtl. 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(1):61–67, 2005.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Muhammad Tausif Irshad, Muhammad Adeel Nisar, Philip Gouverneur, Marion Rapp, and Marcin Grzegorzek. Ai approaches towards prechtl's assessment of general movements: A systematic literature review. *Sensors*, 20(18):5321, 2020.

Namarta Kapil, Bittu Majmudar-Sheth, Alexa Celeste Escapita, and Tara Johnson. Unveiling the immediate impact of prechtl's general movement assessment training on inter-rater reliability and cerebral palsy prediction. *NeuroSci*, 5(3):244–253, 2024.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

Varun Ganjigunte Prakash, Manu Kohli, Aragulla Prasad Prathosh, Monica Juneja, Manushree Gupta, Smitha Sairam, Sadasivan Sitaraman, Anjali Sanjeev Bangalore, John Vijay Sagar Kommu, Lokesh Saini, et al. Video-based real-time assessment and diagnosis of autism spectrum disorder using deep neural networks. *Expert Systems*, 2023.

Kamini Raghuram, Silvia Orlandi, Paige Church, Maureen Luther, Alex Kiss, and Vibhuti Shah. Automated movement analysis to predict cerebral palsy in very preterm infants: an ambispective cohort study. *Children*, 9(6):843, 2022.

Simon Reich, Dajie Zhang, Tomas Kulvicius, Sven Bölte, Karin Nielsen-Saines, Florian B Pokorny, Robert Peharz, Luise Poustka, Florentin Wörgötter, Christa Einspieler, et al. Novel ai driven approach to classify infant motor functions. *Scientific Reports*, 11(1): 9888, 2021.

Rannei Sæther, Ragnhild Støen, Torstein Vik, Toril Fjørtoft, Randi Tynes Vågen, Inger Elisabeth Silberg, Marianne Loennecken, Unn Inger Møinichen, Stian Lydersen, and Lars Adde. A change in temporal organization of fidgety movements during the fidgety movement period is common among high risk infants. *European Journal of Paediatric Neurology*, 20(4):512–517, 2016.

William Thomas Schmidt, Matthew Regan, Michael C Fahey, and Andrew Paplinski. General movement assessment by machine learning: why is it so difficult? *Journal of Medical Artificial Intelligence*, 2(July):15, 2019.

Nelson Silva, Dajie Zhang, Tomas Kulvicius, Alexander Gail, Carla Barreiros, Stefanie Lindstaedt, Marc Kraft, Sven Bölte, Luise Poustka, Karin Nielsen-Saines, et al. The future of general movement assessment: The role of computer vision and machine learning–a scoping review. *Research in developmental disabilities*, 110:103854, 2021.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Ying-Chin Wu, Elisabeth JM Straathof, Kirsten R Heineman, and Mijna Hadders-Algra. Typical general movements at 2 to 4 months: Movement complexity, fidgety movements, and their associations with risk factors and sinda scores. *Early Human Development*, 149: 105135, 2020.

Appendix A. AggPose for Automatic Infant Pose Estimation

To generate automatic keypoint annotations across our video dataset, we employed the *AggPose* framework (Cao et al., 2022), a state-of-the-art vision transformer specifically designed for infant pose estimation. This appendix provides a more detailed overview of the model and our rationale for its use as an automated labelling tool in our pipeline.

Overview and Motivation AggPose (Aggregation Vision Transformer) addresses several core challenges in infant pose estimation, including small body sizes, high variability in posture, and the limited availability of large, annotated datasets. Unlike traditional convolutional or hybrid CNN-transformer models, AggPose uses a fully transformer-based architecture that integrates multi-scale spatial information through a deep-layer aggregation mechanism and cross-resolution fusion via multi-layer perceptrons (MLPs). This design enables the model to learn global and local spatial dependencies without relying on convolutional backbones, making it better suited for fine-grained pose analysis in infants.

The model is initially pre-trained on the COCO dataset and subsequently fine-tuned on a dedicated infant dataset introduced by the authors, which includes over 20,000 manually labelled images and millions of unlabeled frames. This domain-specific adaptation allows AggPose to outperform general-purpose pose estimators in capturing the subtleties of infant movement.

Model Architecture AggPose is structured around a multi-stage transformer pipeline, where each stage processes feature maps at different spatial resolutions using Mix Transformer (MiT) blocks. These consist of overlapped patch embeddings, multi-head self-attention layers, and feedforward modules (Mix-FFNs) that incorporate depth-wise convolution to improve local detail capture.

To facilitate feature integration across resolutions, the model employs an MLP-based fusion mechanism instead of conventional skip connections. This cross-layer aggregation allows for efficient information sharing across spatial scales, improving convergence speed and robustness. Two model variants are available—AggPose-S and AggPose-L—differing in backbone size and depth. According to the original benchmarks, this model achieved 76.4 AP on COCO and 95.0 AP on the infant pose dataset.

Use in Our Study We applied AggPose-L to automatically annotate each video frame in our dataset with 21 infant-specific keypoints—extending the COCO format to include additional clinically relevant joints such as fingers, toes, and the navel. These annotations form the basis for downstream analysis of pose and movement over time. The choice of AggPose was motivated by its strong benchmark performance relative to alternatives like OpenPose, HRNet, and TokenPose, as well as its ability to generalize well to infant data.

Beyond raw accuracy, AggPose offers several practical advantages: reliable keypoint detection across varied poses and lighting conditions; consistent labelling across long sequences; and a clinically meaningful keypoint format. Its transformer-based architecture, paired with efficient aggregation, makes it ideal for large-scale, automated annotation tasks requiring both precision and scalability. An example of the extracted keypoints is shown in Figure 3c.

Appendix B. Further Results

Table 4, Table 5 and Table 6 show the results of the presented classification models, for each labeller individually as well as averaged, grouped by age and feature extraction method used, measured in AUROC, AUPRC and accuracy, respectively. The *AggPose* results are also included. The best performing method per task is marked in bold.

Table 4: **AUROC** of various classifiers for GM classification across different keypoint labellers. The mean and standard deviation over 30 seeds is reported. Best result per model and labeller is bolded.

Keypoint Retrieval	Input Features	RF	LSTM	CNN
Early GM Group				
Label & Track Labeller #1	Coordinates	0.66 ± 0.19	0.64 ± 0.16	0.76 ± 0.19
	Angles	$\mathbf{0.83 \pm 0.13}$	0.66 ± 0.16	$\mathbf{0.82 \pm 0.14}$
	Both	0.79 ± 0.14	$\mathbf{0.67 \pm 0.16}$	0.79 ± 0.17
Label & Track Labeller #2	Coordinates	0.49 ± 0.19	$\mathbf{0.68 \pm 0.17}$	0.65 ± 0.19
	Angles	$\mathbf{0.68 \pm 0.15}$	0.63 ± 0.17	0.61 ± 0.19
	Both	0.57 ± 0.20	0.61 ± 0.16	$\mathbf{0.67 \pm 0.20}$
Label & Track Labeller #3	Coordinates	0.59 ± 0.21	0.62 ± 0.18	0.75 ± 0.20
	Angles	$\mathbf{0.81 \pm 0.12}$	0.56 ± 0.16	$\mathbf{0.81 \pm 0.14}$
	Both	0.73 ± 0.18	$\mathbf{0.66 \pm 0.21}$	0.78 ± 0.17
Label & Track Mean	Coordinates	0.58	**0.65**	0.72
	Angles	**0.77**	0.62	0.75
	Both	0.70	0.65	**0.75**
AggPose	Coordinates	0.48 ± 0.19	0.57 ± 0.19	0.46 ± 0.18
	Angles	$\mathbf{0.51} \pm 0.19$	$\mathbf{0.61} \pm 0.21$	0.49 ± 0.15
	Both	0.48 ± 0.17	0.53 ± 0.21	$\mathbf{0.51} \pm 0.19$
Late GM Group				
Label & Track Labeller #1	Coordinates	0.37 ± 0.25	0.44 ± 0.23	$\mathbf{0.58 \pm 0.23}$
	Angles	$\mathbf{0.51 \pm 0.25}$	$\mathbf{0.61 \pm 0.21}$	0.38 ± 0.22
	Both	0.38 ± 0.24	0.40 ± 0.19	0.54 ± 0.22
Label & Track Labeller #2	Coordinates	0.45 ± 0.24	0.49 ± 0.24	0.51 ± 0.20
	Angles	$\mathbf{0.63 \pm 0.20}$	0.49 ± 0.22	$\mathbf{0.69 \pm 0.22}$
	Both	0.49 ± 0.23	$\mathbf{0.55 \pm 0.18}$	0.59 ± 0.20
Label & Track Labeller #3	Coordinates	0.52 ± 0.27	0.43 ± 0.27	0.62 ± 0.19
	Angles	$\mathbf{0.63 \pm 0.22}$	$\mathbf{0.54 \pm 0.23}$	$\mathbf{0.59 \pm 0.24}$
	Both	0.57 ± 0.18	0.48 ± 0.25	0.54 ± 0.29
Label & Track Mean	Coordinates	0.45	0.45	**0.57**
	Angles	**0.59**	**0.54**	0.55
	Both	0.48	0.47	0.56
AggPose	Coordinates	0.40 ± 0.26	0.54 ± 0.23	0.56 ± 0.24
	Angles	$\mathbf{0.44} \pm 0.25$	$\mathbf{0.66} \pm 0.24$	$\mathbf{0.72} \pm 0.24$
	Both	0.40 ± 0.24	0.65 ± 0.26	0.52 ± 0.29

Table 5: Results. **AUPRC** over 30 seeds is reported. The best result for each model and labeller is bolded.

Keypoint Retrieval	Input Features	RF	LSTM	CNN
Early GM Group				
Label & Track Labeller #1	Coordinates	0.77 ± 0.14	$\mathbf{0.75 \pm 0.13}$	0.86 ± 0.11
	Angles	$\mathbf{0.89 \pm 0.09}$	0.73 ± 0.16	$\mathbf{0.88 \pm 0.10}$
	Both	0.86 ± 0.09	0.74 ± 0.15	0.85 ± 0.13
Label & Track Labeller #2	Coordinates	0.67 ± 0.14	$\mathbf{0.78 \pm 0.14}$	$\mathbf{0.80 \pm 0.12}$
	Angles	$\mathbf{0.77 \pm 0.13}$	0.71 ± 0.17	0.69 ± 0.16
	Both	0.71 ± 0.15	0.71 ± 0.15	0.80 ± 0.14
Label & Track Labeller #3	Coordinates	0.73 ± 0.15	0.70 ± 0.17	0.85 ± 0.13
	Angles	$\mathbf{0.87 \pm 0.09}$	0.67 ± 0.16	0.85 ± 0.14
	Both	0.81 ± 0.13	$\mathbf{0.75 \pm 0.17}$	$\mathbf{0.86 \pm 0.12}$
Label & Track Mean	Coordinates	0.72	**0.74**	**0.84**
	Angles	**0.84**	0.70	0.80
	Both	0.79	0.73	0.84
AggPose	Coordinates	0.63 ± 0.16	0.70 ± 0.15	0.65 ± 0.15
	Angles	$\mathbf{0.68} \pm 0.15$	$\mathbf{0.73} \pm 0.16$	0.65 ± 0.13
	Both	0.63 ± 0.16	0.65 ± 0.16	$\mathbf{0.68} \pm 0.15$
Late GM Group				
Label & Track Labeller #1	Coordinates	0.26 ± 0.14	0.36 ± 0.22	$\mathbf{0.46 \pm 0.23}$
	Angles	$\mathbf{0.39 \pm 0.25}$	$\mathbf{0.47 \pm 0.24}$	0.27 ± 0.17
	Both	0.29 ± 0.19	0.29 ± 0.15	0.40 ± 0.22
Label & Track Labeller #2	Coordinates	0.31 ± 0.18	0.38 ± 0.27	0.48 ± 0.24
	Angles	$\mathbf{0.47 \pm 0.25}$	$\mathbf{0.38 \pm 0.22}$	$\mathbf{0.50 \pm 0.27}$
	Both	0.35 ± 0.18	0.38 ± 0.22	0.46 ± 0.25
Label & Track Labeller #3	Coordinates	0.36 ± 0.19	0.34 ± 0.23	$\mathbf{0.55 \pm 0.21}$
	Angles	$\mathbf{0.49 \pm 0.28}$	$\mathbf{0.42 \pm 0.24}$	0.41 ± 0.25
	Both	0.36 ± 0.18	0.38 ± 0.22	0.43 ± 0.27
Label & Track Mean	Coordinates	0.31	0.36	**0.50**
	Angles	**0.45**	**0.42**	0.39
	Both	0.33	0.35	0.43
AggPose	Coordinates	0.36 ± 0.21	0.50 ± 0.24	0.50 ± 0.25
	Angles	$\mathbf{0.36} \pm 0.19$	0.58 ± 0.27	$\mathbf{0.65} \pm 0.28$
	Both	0.33 ± 0.18	$\mathbf{0.59} \pm 0.28$	0.45 ± 0.28

Table 6: Results. **Accuracy** over 30 seeds is reported. The best result for each model and labeller is bolded.

Keypoint Retrieval	Input Features	RF	LSTM	CNN
Early GM Group				
Label & Track Labeller #1	Coordinates	0.60 ± 0.16	0.59 ± 0.15	0.69 ± 0.15
	Angles	$\mathbf{0.73 \pm 0.13}$	0.60 ± 0.14	$\mathbf{0.70 \pm 0.16}$
	Both	0.66 ± 0.15	$\mathbf{0.64 \pm 0.16}$	0.68 ± 0.12
Label & Track Labeller #2	Coordinates	0.45 ± 0.17	$\mathbf{0.60 \pm 0.15}$	0.62 ± 0.13
	Angles	$\mathbf{0.65 \pm 0.14}$	0.56 ± 0.14	0.56 ± 0.17
	Both	0.53 ± 0.18	0.59 ± 0.16	$\mathbf{0.63 \pm 0.15}$
Label & Track Labeller #3	Coordinates	0.55 ± 0.18	0.61 ± 0.14	0.68 ± 0.14
	Angles	$\mathbf{0.70 \pm 0.11}$	0.54 ± 0.15	$\mathbf{0.75 \pm 0.12}$
	Both	0.62 ± 0.17	$\mathbf{0.65 \pm 0.17}$	0.71 ± 0.14
Label & Track Mean	Coordinates	0.54	0.60	0.66
	Angles	**0.69**	0.57	0.67
	Both	0.60	**0.63**	**0.67**
AggPose	Coordinates	0.47 ± 0.13	0.54 ± 0.16	0.47 ± 0.15
	Angles	$\mathbf{0.50} \pm 0.17$	$\mathbf{0.59} \pm 0.14$	0.47 ± 0.11
	Both	0.47 ± 0.15	0.53 ± 0.20	$\mathbf{0.49} \pm 0.14$
Late GM Group				
Label & Track Labeller #1	Coordinates	0.58 ± 0.14	0.58 ± 0.11	0.61 ± 0.16
	Angles	$\mathbf{0.63 \pm 0.17}$	$\mathbf{0.68 \pm 0.13}$	0.57 ± 0.13
	Both	0.61 ± 0.16	0.58 ± 0.16	$\mathbf{0.62 \pm 0.15}$
Label & Track Labeller #2	Coordinates	0.64 ± 0.13	0.59 ± 0.15	0.56 ± 0.21
	Angles	$\mathbf{0.68 \pm 0.08}$	0.63 ± 0.14	$\mathbf{0.64 \pm 0.13}$
	Both	0.64 ± 0.13	$\mathbf{0.64 \pm 0.14}$	0.61 ± 0.16
Label & Track Labeller #3	Coordinates	0.67 ± 0.14	0.60 ± 0.15	0.61 ± 0.19
	Angles	$\mathbf{0.70 \pm 0.08}$	0.62 ± 0.16	0.64 ± 0.12
	Both	0.66 ± 0.09	$\mathbf{0.65 \pm 0.17}$	$\mathbf{0.64 \pm 0.13}$
Label & Track Mean	Coordinates	0.63	0.59	0.60
	Angles	**0.67**	**0.64**	0.62
	Both	0.64	0.62	**0.62**
AggPose	Coordinates	0.54 ± 0.17	0.60 ± 0.17	0.62 ± 0.18
	Angles	$\mathbf{0.62} \pm 0.11$	$\mathbf{0.71} \pm 0.15$	$\mathbf{0.70} \pm 0.18$
	Both	0.59 ± 0.14	0.61 ± 0.18	0.61 ± 0.19

Appendix C. Label & Track vs. AggPose

Table 7 shows the difference in performance between the *Label & Track* and *AggPose* methods for keypoint tracking, statistically significant ($p < 0.05$) differences are bolded. For comparisons where one of the two methods is not normally distributed (according to the Shapiro-Wilk test), the Mann-Whitney U Test was used, otherwise a t-test was used. Table 8 shows the calculated p values.

Table 7: Difference in performance of the Label & Track method (mean over all labellers) vs. AggPose, bolded differences are statistically significant ($p < 0.05$), p-values are given in the following table. Values marked with an asterisk were calculated using a t-test, all other values used the Mann-Whitney U test.

Metric	Input Features	Early GM Group			Late GM Group		
		RF	LSTM	CNN	RF	LSTM	CNN
AUROC	Coord.	**0.1005***	**0.1226***	**0.2578**	0.0463	−0.0889∗	0.0071
	Angles	**0.2641**	0.0075∗	**0.2546**	**0.1511**	**-0.1146***	**-0.1625**
	Both	**0.2122**	**0.1091***	**0.2383**	0.0735∗	**-0.1737***	0.0342∗
AUPRC	Coord.	**0.0890***	**0.0954**	**0.1865**	−0.0452	**-0.1412**	0.0008
	Angles	**0.1623**	−0.0290	**0.1500**	0.0878	**-0.1529**	**-0.2629**
	Both	**0.1587**	**0.0857**	**0.1591**	−0.0005	**-0.2397**	−0.0207
Accuracy	Coord.	0.0671∗	**0.0540**	**0.1980**	**0.0939**	−0.0122∗	−0.0277
	Angles	**0.1912***	−0.0235∗	**0.2000***	**0.0504**	**-0.0693***	**-0.0876**
	Both	**0.1271**	**0.0987***	**0.1863**	0.0426	0.0143∗	0.0098

Table 8: P values for relative performance of the Label & Track method vs. AggPose, bolded differences are statistically significant ($p < 0.05$). Values marked with an asterisk were calculated using a t-test, all other values used the Mann-Whitney U test.

Metric	Input Features	Early GM Group			Late GM Group		
		RF	**LSTM**	**CNN**	**RF**	**LSTM**	**CNN**
AUROC	Coord.	**0.0213***	**0.0016***	**0.0000**	0.3728	0.0908∗	0.9536
	Angles	**0.0000**	0.8431∗	**0.0000**	**0.0075**	**0.0192***	**0.0051**
	Both	**0.0000**	**0.0069***	**0.0000**	0.1381∗	**0.0004***	0.5183∗
AUPRC	Coord.	**0.0070***	**0.0040**	**0.0000**	0.3200	**0.0030**	0.8771
	Angles	**0.0000**	0.4397	**0.0000**	0.1813	**0.0040**	**0.0000**
	Both	**0.0000**	**0.0107**	**0.0000**	1.0000	**0.0000**	0.7548
Accuracy	Coord.	0.0547∗	**0.0405**	**0.0000**	**0.0048**	0.6917∗	0.8148
	Angles	**0.0000***	0.4485∗	**0.0000***	**0.0325**	**0.0248***	**0.0071**
	Both	**0.0008**	**0.0072***	**0.0000**	0.0742	0.6787∗	0.6325