Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France – August, 2025
Paper No. ICBES 126
DOI: 10.11159/icbes25.126

Training General Movements Classifiers with Global Labels Offers Insights on Sub-movement Quality

Manpreet Kaur[1], Hamid Abbasi[2], Sîan A. Williams[3], Malcolm R. Battin[4], Thor F. Besier[2], Angus J.C. McMorland[1]

[1]Department of Exercise Sciences
The University of Auckland, Auckland, 1023, New Zealand
manpreet.kaur@auckland.ac.nz; a.mcmorland@auckland.ac.nz
[2]Auckland Bioengineering Institute
The University of Auckland, Auckland, 1010, New Zealand
h.abbasi@auckland.ac.nz; t.besier@auckland.ac.nz
[3]Liggins Institute
The University of Auckland, Auckland, 1023, New Zealand
sian.williams@auckland.ac.nz
[4]Auckland City Hospital
Auckland, 1023, New Zealand
malcolmb@adhb.govt.nz

Abstract – Abnormal or absent General Movements (GMs) during the fidgety period (9–20 weeks post-term) are strong early indicators of neurological disorders, including cerebral palsy (CP). The General Movements Assessment (GMA) is the clinical gold standard for evaluating GMs, but its reliance on expert assessment limits accessibility and scalability. Machine Learning (ML)-based models offer a promising alternative in automating movement classification; however, existing approaches to training these systems either require extensive manual annotation of short segments of infant movements (or snippets) or classify entire videos without capturing movement-level details. This study addresses these limitations by demonstrating that a ML classifier trained with video-level (per infant) labels can accurately classify whole videos of infant movements and provide useful information about movement snippets. We trained and evaluated several models, including SVM, LSTM, 1D-CNN, and Vision Transformer (ViT), using time-series representations of infant movements. The best-performing model, a 1D-CNN, achieved 100% accuracy in video-level classification and 87.5% accuracy in snippet-level (i.e., movement-level) classification of previously unseen data, using 2D coordinates of 24 body landmarks and 12 joint angle features. Additionally, we examined whether the feature space of videos labelled as normal and abnormal shows overlap, using Independent Component Analysis (ICA) and cosine-similarity between 1D-CNN abstractions. Our findings align with clinical observations, indicating that short movement segments from infants labelled as abnormal share characteristics with those from infants with normal GMs, impacting classification performance. Overall, this work provides insights useful for working towards fully automated GMA analysis capable of providing both movement- and video-level assessment, which will enhance early prediction of neurodevelopmental abnormalities with improved scalability.

Keywords**:** Cerebral Palsy (CP), General Movements Assessment (GMA), Infant Motion Tracking, Infant Movement Classification, Machine Learning

1. Introduction

Spontaneous general movements (GMs), particularly fidgety movements (FMs) observed between 9 and 20 weeks post-term, are strong indicators of an infant's neurodevelopmental status [1]. The General Movements Assessment (GMA) is a well-established clinical method for evaluating GMs and categorizing them as normal, abnormal, or absent to provide critical insights into early neurological function [2]. During the fidgety period, normal GMs are characterized by small, circular, and fluent movements of the neck, trunk, and limbs, with moderate speed and variable acceleration. In contrast, abnormal GMs, such as cramped synchronized or absent FMs, are strongly associated with neurological disorders, including cerebral palsy (CP) [3]. Early identification of these abnormalities allows for early interventions when neuroplasticity is at its peak,

maximising therapeutic effectiveness [4]. However, GMA's reliance on expert visual assessment makes it time-intensive, subjective, and difficult to scale to larger populations. The need for certified professionals further restricts accessibility, potentially delaying early diagnosis and intervention.

To address the challenges associated with manual GMA, several studies [5, 6] have explored machine learning (ML) models trained on annotated movement snippets (labelled for the presence or absence of FMs) segmented from video recordings to automate the classification of normal and abnormal movements. Other approaches [7, 8] have classified infant movements as typical or not by extracting features from the entire video without segmenting them. Snippet-level (or movement-level) classification requires extensive manual annotation of hundreds of snippets to generate ground-truth labels, making the process time-intensive. Furthermore, infant movement classification extends beyond a binary normal/abnormal distinction, as mildly abnormal infants can have periods with similar movement patterns to typically developing infants, also exhibiting complexity and variation in a subset of their movements [3, 9, 10]. This movement-level insight is not provided by video-level classifiers; instead, a single label is assigned to the infant for the entire video based on a general analysis of overall movements.

This study addresses the limitations described above. Firstly, we developed an ML-based classifier for classifying whole videos (or infants) into normal or abnormal categories trained using global video labels. Secondly, we show that the resultant classifier, despite being trained on whole-video labels, provides useful information about movement snippets as well: 'misclassifications' of individual snippets from infants producing abnormal GMs within this analysis actually correspond to movement periods exhibiting characteristics that are similar to normal movements. Independent Component Analysis (ICA) of our input kinematic features supports the idea that infants producing abnormal GMs share some movement characteristics with normal movements.

2. Data Acquisition

2.1. Ethics

All procedures were approved by the Auckland Health Research Ethics Committee (000146). Parents/caregivers were fully informed about the study's purpose, including the filming procedure and methods, and provided written consent for their child's participation.

2.1. Experimental Procedure

The study involved 13 full-term infants (8 normal, 5 abnormal), recorded during the fidgety period, each contributing a 2–3-minute video. The cohort had a mean and standard deviation, in brackets, gestational age of 218.46 (36.61) days at birth, birth weight of 1619.23 (947.66) g, and corrected age of 89.07 (11.21) days at the time of recording. Videos were captured from a top-down view using a standard iPad (1920×1080 resolution, 29.97 frames per second), with infants dressed in diapers. Videos were cropped using Adobe Premiere Pro 2021 to keep infants centered in the frame.

All the experiments were performed on the New Zealand eScience Infrastructure, using four A100 GPUs, each with 40 GB of memory. ML classifiers were trained using TensorFlow v2.6.2, CUDA v11.6, Python v3.6.8, and Scikit-learn v1.6.

3. Methods

In this work, we employed a pose estimation tool for estimating 2D landmarks for each frame, followed by a custom post-processing algorithm to correct mislabelled locations. Then we computed kinematic feature time series and trained ML classifiers. All these components are described in detail below.

3.1. Pose Estimation and Data Pre-processing

We utilized DeepLabCut (DLC) [11], an open-source, markerless pose estimation model, to predict 2D coordinates of 24 landmarks, as introduced in our previous work [12]. Compared to other models such as OpenPose [13] and MediaPipe [14], DLC provides greater flexibility, allowing users to track customised or specific body landmarks rather than relying on predefined ones.

The estimated poses from DLC contained some incorrect landmark predictions. To correct these, we performed the following pre-processing steps: (1) *Filtering incorrect predictions*: an in-house built algorithm was employed to filter out incorrectly predicted landmarks based on skeleton geometry and trajectory smoothness; (2) *Data imputation*: missing landmarks were imputed using modified Akima interpolation [15]; and (3) *Manual correction*: interpolated trajectories were visually inspected by authors (MK) using video overlays, and interpolation errors occurring due to large gaps were manually corrected.

Despite working with a small dataset, we preserved the real-world variability, including diverse environmental conditions, body sizes, and camera distances, through video selection and by intentionally avoiding adjustments that could reduce these effects. This approach enhances the classifier's ability to generalize to unseen data [16]. The only normalization applied was a translation of all data points that aligned each infant's sternum to (0,0).

For training all the classifiers, we used three types of feature sets: (1) 2D coordinates from 24 landmarks; (2) 12 joint angles (shoulders, elbows, wrists, hips, knees and ankles for both sides); and (3) combination of both (1) and (2).

3.2. Train/Test Data and Annotations

All 13 videos (8 normal, 5 abnormal) were clinically assessed by a qualified clinician (SW) following GMA guidelines, assigning each infant a global label of normal (0) or abnormal (1) to provide ground truth.

The cohort was divided into two sets: Set 1, with 10 videos (6 normal, 4 abnormal), for internal validation and Set 2, with 3 videos (2 normal, 1 abnormal), for external validation. All the videos were segmented into 10 s (300-frame) snippets, as shown in Fig. 1 [17]. The selected window length was recommended by GMA-certified clinicians.

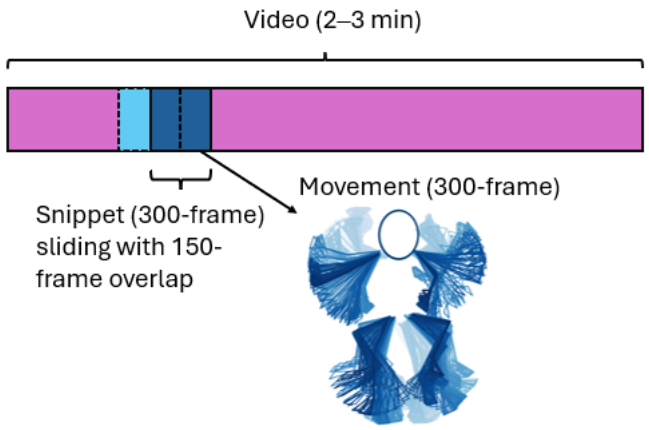

Fig. 1: An example of 300-frame snippet sliding with 150-frame overlap and corresponding movement trajectories [17].

To examine the impact of training sample size and movement redundancy from overlapping in training snippets on classifier performance, Set 1 was used to generate datasets with different window overlap values (299, 290, 270, 250, 200, 150, and 100 frames). Snippets were assigned the global label of their respective video. The largest dataset was obtained with an overlap of 299 frames, containing 11122 snippets from normal videos, and 3209 snippets from abnormal videos. The smallest dataset, with a 100-frame overlap, contained 58 normal labelled and 19 abnormal labelled snippets.

Internal validation was performed using 5-fold cross-validation, ensuring balanced class distribution across rounds. To ensure reproducibility and consistency of the data split (and their associated indices) across all runs, the '*random_state'* of the 'train_test_split' function was set at 42. In each round, one fold served as the unseen test set of the internal step, while the remaining four were split into a training set (90%) and a validation set (10%; 5% from each class). This validation set was used to monitor performance and prevent overfitting via early stopping and halting training if the validation loss remained unchanged for 10 consecutive epochs. The trained classifier was then evaluated on the test set using the best parameter values. Whereas the internal validation k-fold process assessed performance on previously unseen snippets, it did

not exclude testing on other snippets from the same videos as those in the training data. External validation was performed to test the generalizability of the classifier with the highest accuracy from internal validation on completely unseen videos.

Data augmentation was performed by flipping the right-left 2D coordinates and joint angles, while noise augmentation was avoided to preserve fine-grained GMs crucial for accurate assessment. Before the classification, all features were standardized using z-score normalization, adjusting the mean to zero and the standard deviation to one, facilitating faster convergence and preventing feature dominance.

3.3. Supervised Machine Learning Classification

We explored both traditional and advanced ML models for infant movement classification, including Support Vector Machines (SVM), Long Short-Term Memory (LSTM) networks, 1D Convolutional Neural Networks (1D-CNN), and a modified Vision Transformer (ViT). Given that classifier architecture significantly impacts generalization to unseen data, we optimized model performance by exploring different architectures and hyperparameter configurations.

SVM, well-suited for high-dimensional data, was implemented with two kernel types: Radial Basis Function (RBF) and third-degree polynomial. LSTM and 1D-CNN models, designed to capture temporal dependencies and complex spatial patterns, were evaluated using both two- and three-layer configurations. For LSTM, we tested neuron configurations of (64, 32) and (128, 64, 32), while for 1D-CNN, we used filter sizes (32, 64) and (32, 64, 128), followed by dense layers with 32 and 64 neurons, respectively. A final sigmoid-activated neuron was incorporated into both LSTM and 1D-CNN for binary classification.

The ViT, known for its self-attention mechanism and effectiveness in image classification, was adapted for infant movement time series data. The original ViT architecture was modified by adjusting its input layers to process time series data while retaining its encoder structure. Positional encoding was applied to each frame within the snippet to preserve temporal relationships. The model was trained from scratch, as the original ViT was pre-trained on image data. Given the medium size of the dataset and to avoid potential overfitting associated with utilizing complex ViT architectures on these data, we used a ViT with eight attention heads and two depth configurations (2 and 4 encoder layers), followed by a fully connected layer (128 neurons) before the classification neuron.

A dropout rate of 0.3 was applied across all models. Training was conducted using the ADAM optimizer with a batch size of 16 for up to 250 epochs.

For video-level (i.e., infant-level) classification, a threshold of 0.6 was applied: if more than 60% of an infant's snippets were classified as normal, the infant was assigned a normal label; otherwise, abnormal. This threshold ensures that the majority of an infant's movements are classified within a specific category while maintaining the balance in cases where overlapping movement patterns may lead to the misclassification of certain snippets.

3.4. Quantification of Overlapping Movement Characteristics

To quantitatively assess the clinically observed overlap between GMA-labelled normal and abnormal infants, two approaches were used. Firstly, Independent Component Analysis (ICA) was applied to the Set 1 dataset with 150-frame overlap, decomposing movement features into two independent components for further analysis. Secondly, feature vectors from the final non-classification layer of the best-performing 1D-CNN were compared between different snippets using the cosine similarity metric.

4. Results

4.1 Internal Validation Classification Performance

Fig. 2 illustrates the classification performance of the best-performing classifiers of each type, tested with different architectures, at various overlap levels. All classifiers achieved 100% snippet-level accuracy when trained on datasets with 299-, 290-, and 270-frame overlaps. This was potentially due to the redundancy of repeated movement patterns and the larger training sample size, which facilitated better learning but may have led to overfitting. As the overlap decreased to 150 or 100 frames, variability between consecutive snippets increased, leading to a decline in snippet-level classification accuracy. We

subsequently focused on results obtained with a 150-frame overlap for further analysis, which aligns with clinical recommendations while balancing movement variability and redundancy.

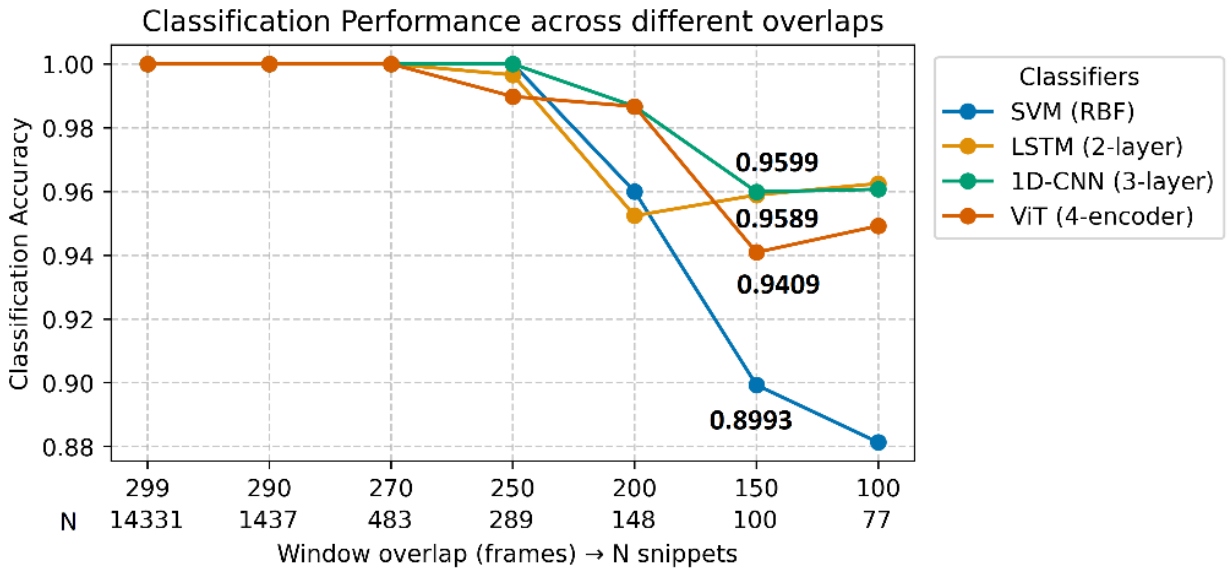

Fig. 2: Snippet-level classification performance of best-performing classifiers across different overlaps. X axis labels show window overlap (top) and number of snippets (train + test; bottom). Numbers in bold indicate accuracy of each classifier for 150-frames overlap.

Notably, the three-layer 1D-CNN outperformed all other classifiers, achieving 95.99% accuracy in movement classification. In contrast, SVM demonstrated the lowest accuracy (89.93%). ViT, despite being trained from scratch, achieved 94.09% accuracy—lower than 1D-CNN and LSTM, but demonstrating the potential of vision transformers for time-series-based infant movement classification.

4.2 External Validation Classification Performance

We tested the performance of the 1D-CNN classifiers developed above on completely unseen videos. The best-performing classifier achieved snippet-level accuracy of 87.5% (others: 62.5%, 41.7%, 4.2%).

The best classifier successfully predicted correct labels for snippets from normal infants, while for the infant with abnormal GMs, 6 out of 9 snippets were classified as abnormal and 3 as normal. These 'misclassifications' likely reflect overlap in movement characteristics between normal and abnormal infants: healthy infants do not always perform GMs, and infants classified with abnormal GMs do perform some normal GM-like movements. Nevertheless, since the proportion of abnormal-classified snippets exceeded the 0.6 threshold, the overall infant label was correctly assigned as abnormal, resulting in 100% accuracy for infant-level classification on previously unseen data.

4.3 Quantification of Overlapping Movement Characteristics

To examine the overlap in movement characteristics, we applied ICA to Set 1 with 150-frame overlap using a 5-fold approach, evaluating all three feature sets (as explained above). ICA, which is effective in extracting independent and uncorrelated features, resulted in the best feature separation when applied to joint angles compared to the other two feature sets. Fig. 3(A) shows that the first independent component for each fold demonstrated one narrow peak for normal videos. The equivalent distribution from videos of abnormal movements was wider, containing a main peak distinct from the normal data, as well as an overlapping region, potentially corresponding to similar movement characteristics.

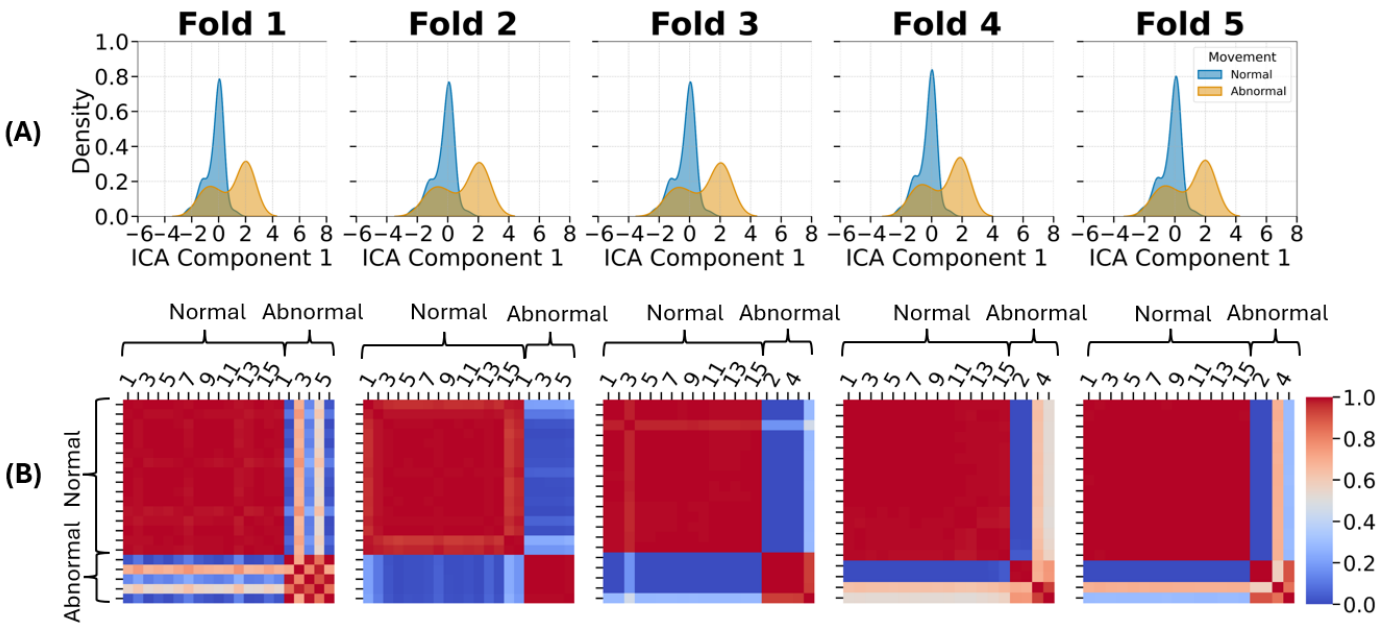

Fig. 3: Overlapping movement feature patterns from Set 1(A) after applying ICA (B) heatmap of cosine similarity between normal and abnormal features learned by 1D-CNN, where both x- and y-axes represent the same snippet labels.

Additionally, to investigate the impact of feature overlap on classification performance, we compared 1D-CNN abstracted features using the cosine similarity between normal and abnormal snippet features, plotted as heatmaps in Fig. 3(B). This revealed that some abnormal snippets, classified as normal by the snippet-level classification analysis, specifically in folds 1 (Abnormal-2 and -4), 4 (Abnormal-3), and 5 (Abnormal-3), were more similar to normal movements than abnormal ones, suggesting shared movement patterns across the groups. These 'misclassifications' indicate that while the model learned distinguishing features, it also captured underlying similarities, aligning with clinical observations of movement variability beyond a strict normal/abnormal distinction.

5. Conclusion

In this work, we developed a classifier capable of accurately distinguishing between normal and abnormal infant movements at both the snippet and video levels. Our best-performing model, a 1D-CNN, achieved 87.5% accuracy in movement classification and 100% accuracy in video-level classification on previously unseen data, demonstrating its robustness in recognizing movement patterns at a clinically recommended resolution without requiring snippet-level annotations. Our ICA analysis supports existing literature, indicating potential overlapping patterns of activity in the developing brains of normal and abnormal GMA-classified infants, highlighting shared movement characteristics despite clinical distinctions.

Acknowledgements

The research was supported by the Friedlander Foundation (Grant: 3720759).

References

[1] H. F. Prechtl, "State of the art of a new functional assessment of the young nervous system. An early predictor of cerebral palsy," *Early Hum. Dev.,* vol. 50, *(1),* pp. 1-11, 1997.

[2] H. F. R. Prechtl, "General movement assessment as a method of developmental neurology: new paradigms and their consequences. The 1999 Ronnie MacKeith Lecture," *Developmental Medicine and Child Neurology,* vol. 43, *(12),* pp. 836-842, 2001.
[3] C. Einspieler and H. F. Prechtl, "Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system," *Ment. Retard. Dev. Disabil. Res. Rev.,* vol. 11, *(1),* pp. 61-67, 2005.
[4] M. Hadders-Algra, "Early diagnostics and early intervention in neurodevelopmental disorders—age-dependent challenges and opportunities," *Journal of Clinical Medicine,* vol. 10, *(4),* pp. 861, 2021.
[5] Reich, S., Zhang, D., Kulvicius, T., Bölte, S., Nielsen-Saines, K., Pokorny, F.B., Peharz, R., Poustka, L., Wörgötter, F., Einspieler, C. and Marschik, P.B., "Novel AI driven approach to classify infant motor functions," *Nature,* 2021.
[6] Q. Gao, S. Yao, Y. Tian, C. Zhang, T. Zhao, D. Wu, G. Yu and H. Lu, "Automating General Movements Assessment with quantitative deep learning to facilitate early screening of cerebral palsy," *Nature Communications,* vol. 14, *(1),* pp. 8294, 2023.
[7] B. Nguyen-Thai, V. Le, C. Morgan, N. Badawi, T. Tran and S. Venkatesh, "A Spatio-Temporal Attention-Based Model for Infant Movement Assessment From Videos," *IEEE J. Biomed. Health Inform.,* vol. 25, *(10),* pp. 3911, -10, 2021.
[8] K. D. Mccay, E. S. L. Ho, H. P. H. Shum, G. Fehringer, C. Marcroft and N. D. Embleton, "Abnormal Infant Movements Classification With Deep Learning on Pose-Based Features," *IEEE Access,* vol. 8, pp. 51582, 2020.
[9] M. Hadders-Algra, "General movements: a window for early identification of children at high risk for developmental disorders," *The Journal of Pediatrics,* vol. 145, *(2),* pp. S12, -08, 2004.
[10] I. A. Solopova, V. A. Selionov, I. Dolinskaya and E. S. Keshishian, "General Movements as a Factor Reflecting the Normal or Impaired Motor Development in Infants," *Hum. Physiol.,* vol. 46, pp. 432-442, 2020.
[11] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis and M. Bethge, "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning," *Nat Neurosci,* vol. 21, *(9),* pp. 1281, -08-20, 2018.
[12] M. Kaur, H. Abbasi, S. A. Williams, M. R. Battin, T. F. Besier and A. J. McMorland, "Enhanced markerless tracking of infant general movements in standard videos through lightning pose compared to DeepLabCut," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 2024, pp. 1-4.
[13] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, pp. 7291-7299.
[14] C. Lugaresi, J. Tang, H. Nash, C. Mcclanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, M. Grundmann and G. Research, "MediaPipe: A framework for building perception pipelines," June. 2019.
[15] H. Akima, "A new method of interpolation and smooth curve fitting based on local procedures," *Journal of the ACM (JACM),* vol. 17, *(4),* pp. 589-602, 1970.
[16] Z. Gong, P. Zhong and W. Hu, "Diversity in machine learning," *IEEE Access,* vol. 7, pp. 64323-64350, 2019.
[17] M. Kaur, H. Abbasi, S. A. Williams, M. R. Battin, T. F. Besier and A. J. McMorland, "An example of 300-frame snippet sliding with 150-frame overlap and corresponding movement trajectories, *"https://Figshare.Com/,* 2025. Available: https://figshare.com/s/80a4638920f70c3b94e0.