A Pose-Based Feature Fusion and Classification Framework for the Early Prediction of Cerebral Palsy in Infants

Kevin D. McCay, Pengpeng Hu, Hubert P. H. Shum, *Senior Member, IEEE*, Wai Lok Woo, *Senior Member, IEEE*, Claire Marcroft, Nicholas D. Embleton, Adrian Munteanu, and Edmond S. L. Ho

***Abstract*— The early diagnosis of cerebral palsy is an area which has recently seen significant multi-disciplinary research. Diagnostic tools such as the General Movements Assessment (GMA), have produced some very promising results. However, the prospect of automating these processes may improve accessibility of the assessment and also enhance the understanding of movement development of infants. Previous works have established the viability of using pose-based features extracted from RGB video sequences to undertake classification of infant body movements based upon the GMA. In this paper, we propose a series of new and improved features, and a feature fusion pipeline for this classification task. We also introduce the RVI-38 dataset, a series of videos captured as part of routine clinical care. By utilising this challenging dataset we establish the robustness of several motion features for classification, subsequently informing the design of our proposed feature fusion framework based upon the GMA. We evaluate our proposed framework's classification performance using both the RVI-38 dataset and the publicly available MINI-RGBD dataset. We also implement several other methods from the literature for direct comparison using these two independent datasets. Our experimental results and feature analysis show that our proposed pose-based method performs well across both datasets. The proposed features afford us the opportunity to include finer detail than previous methods, and further model GMA specific body movements. These new features also allow us to take advantage of additional body-part specific information as a means of improving the overall classification performance, whilst retaining GMA relevant, interpretable, and shareable features.**

***Index Terms*— Cerebral palsy, early diagnosis, explainable AI, general movements assessment, machine learning, motion analysis, skeletal pose estimation.**

Manuscript received June 11, 2021; revised November 17, 2021; accepted December 21, 2021. Date of publication December 23, 2021; date of current version January 28, 2022. This work was supported in part by the Royal Society under Grant IES\R2\181024 and Grant IES\R1\191147. *(Corresponding author: Edmond S. L. Ho.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the host organisation, the Research Ethics Committee (REC), the Health Research Authority (HRA), and Health and Care Research Wales (HCRW), under Approval Nos. 19/LO/0606 and under IRAS Project ID: 252317.

Kevin D. McCay, Wai Lok Woo, and Edmond S. L. Ho are with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. (e-mail: e.ho@northumbria.ac.uk).

Pengpeng Hu and Adrian Munteanu are with the Department of Electronics and Informatics, Vrije Universiteit Brussel, 1050 Ixelles, Belgium.

Hubert P. H. Shum is with the Department of Computer Science, Durham University, Durham DH1 3LE, U.K.

Claire Marcroft and Nicholas D. Embleton are with the Newcastle upon Tyne Hospitals Trust, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, U.K.

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2021.3138185, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2021.3138185

I. Introduction

CEREBRAL palsy (CP) is the collective term given to a group lifelong neurological conditions caused by non-progressive damage to the brain [5], occurring before, during, or shortly after birth [49]. CP typically affects movement, muscle tone, posture and co-ordination, but can also cause difficulties with swallowing, speech-articulation, hearing, vision, and can impact upon an individual's ability to learn new skills [17].

CP is the most prevalent physical disability found in children, with 2.11 diagnoses per 1000 live births [41]. There is also an increased prevalence of CP in infants born prematurely, with 32.4 diagnoses per 1000 infants born very preterm (28-32 weeks gestation), and 70.6 diagnoses per 1000 infants born extremely preterm (<28 weeks gestation) [47]. As the degree of prematurity increases so to does the likelihood of severe disability, with one-in three of those surviving at 22 weeks gestation being severely disabled [14], [28]. As such, the early diagnosis of CP is an ongoing area of multidisciplinary research, as it has the potential to allow for early intervention clinical care. Early interventions look to optimise the neuroplasticity of the developing infant brain, thereby inhibiting the impact of impairment and subsequently reducing the likelihood of the child fully developing associative conditions [24]. However, early diagnosis can be difficult and time consuming [31].

Manual examinations, such as the General Movements Assessment (GMA) [16], have been developed to identify the emerging signs of CP. Studies suggest that the GMA

compares favourably with other methods, such as cranial ultrasound and neurological examination, by producing more consistent and reliable individual results [6], with the GMA being identified as the leading method of predicting later CP [26]. The GMA is typically carried out by evaluating the quality of body movements at a specific window in an infant's development (Section II-A); however, the application of these assessments can be challenging, chiefly due to the availability of appropriately skilled clinicians [31].

There is increasing motivation to utilise technology to aid with clinical decision making, helping with logistical constraints, increasing the predictive accuracy and targeting early intervention [39]. As such, several works [1], [33]–[35], [40], [53] have proposed automated solutions to help address the challenges faced in the diagnosis of CP. The proposed methods typically make use of machine learning frameworks to automatically assess infants based upon the movement patterns associated with the GMA. These methods suggest that a machine learning framework could substantiate the decision making process, allowing for intuitive, quantitative, cost-effective, evidence-based evaluation [29], and also provide a means of fully remote diagnostic assessment.

However, it is also clear that the interpretability of the model must be considered, particularly when using machine-learning based approaches in the medical domain. Machine learning models are often seen as 'black boxes', in which the underlying structures can be difficult to comprehend. Consequently, clarity is required as to why a system arrives at a decision, and as such the mechanisms behind classification frameworks have to be transparent, understandable and explainable [21]. We therefore suggest that by using pose data in combination with histogram representations, we retain human interpretability throughout the classification pipeline. As such, we propose a feature-extraction, feature-fusion, and classification framework, which extracts several new GMA relevant features using pose data generated from standard 2D RGB video. By using this approach we aim to ensure that the extracted features are not only closely mapped to the assessment criteria specified in the GMA, but are also able to be understood by clinicians as a means of retaining predictive accountability and explainability. This approach also provides opportunities for remote data analysis without the need for specialist equipment, making wider clinical diagnosis feasible.

To evaluate our system and to inform the design of our proposed features, we subsequently collected a new video dataset (RVI-38), which reflects the complexity associated with video data gathered in real-world clinical settings for use in the GMA. By utilising the challenging RVI-38 dataset, we have developed and evaluated combinations of several different supplementary pose-based features. In doing so, we have produced a feature-set capable of generalising well across datasets of varying size, quality, and duration, whilst simultaneously dealing with superfluous information present in shot. We compare our method with several methods from the literature using both the RVI-38 dataset and the MINI-RGBD [20] dataset to evaluate the comparative robustness of our features. The source code and dataset used in this paper are available at (https://github.com/edmondslho/Pose-basedCerebralPalsyPrediction).

The contributions of this work can be summarised as follows:

- A series of new and enhanced pose-based interpretable features, extracted from 2D video sequences, and based upon specific criteria set out in the GMA.
- Experimental re-implementation, comparative evaluation, and discussion of several prominent previous methods, along with our proposed method, are undertaken using shared datasets, for unbiased assessment and the generation of a new benchmark.
- The introduction of a challenging new video dataset, composed of real-world patient data. This GMA specific dataset informs our proposed feature design and comparative analysis. Additionally, given the difficulty in acquiring data in this sensitive area, we also make the extracted pose dataset and associated GMA labels available to the community.
- Analysis of a new automated feature extraction, fusion, and classification pipeline, for the prediction of cerebral palsy based upon the GMA. We also make this framework available to the community to further encourage research in this field.

II. Related Works

In this section, we discuss the GMA as well as relevant studies which have attempted to automate this diagnostic process using computer vision based approaches. We provide an overview of the associated feature extraction processes, with a view to including several of the discussed methods as comparative evaluation baselines.

A. General Movements Assessment

The GMA is a non-invasive and non-intrusive physical examination of infants for the early detection of neurological anomalies associated with CP. General movements (GMs) are body movements which are spontaneous and variable, and form part of a distinct pattern of movements, called Fidgety Movements (FMs), which can be observed from early fetal life through to around 20 weeks post-term [16].

In a typically developing infant, FMs wax and wane in intensity, speed, frequency, amplitude, and range of motion, with notable fluctuations to the rotation, orientation, and displacement around the limb axes [42]. Conversely, abnormal GMs are identified by the absence of FMs, with a lack of duration, variability, and complexity throughout the movement sequence [48]. FMs are seen as a reliable indicator of brain dysfunction and as such, the presence of these abnormal GM patterns is seen as a strong predictor that the infant will receive a confirmed diagnosis of CP in later life [15].

The ability to provide an early diagnosis of CP is considered key, as early intervention care can contribute towards improved motor-function ability due to the application of neuro-developmental treatments, which optimise the "brain plasticity" found in developing infants [37]. As such, the earlier a confirmed diagnosis can be made, the better the outlook for the infant's long term development [58]. However,

whilst diagnostic tools such as the GMA have been proven to provide accurate prediction of CP, the training required to become a suitably proficient assessor is significant.

Motivated by this, an initial study by Adde *et al.* [1], proposed a system in which computer vision techniques were employed to assess the feasibility of undertaking automated GMA. In this study, a representation of the associated motion was created by generating a "motion image", which quantified the amount of infant movement present in each video through frame differencing. The generated motion images were then used to generate features for binary classification of normal or abnormal infant movements. However the use of difference images is subject to several limitations, such as sensitivity to camera movement, issues with self-occlusion, strict background pre-processing, and a lack of information about the speed and direction of objects moving in frame [43], [53]. As such, several studies have subsequently attempted to further evaluate the viability of automated GMA through more advanced computer vision-based techniques, such as optical flow.

B. Automated Optical Flow-Based Methods

In optical flow based methods, the motion is represented by a displacement vector field of the pixels between consecutive frames of an image sequence, rather than simply calculating the gross quantity of motion present in the frame and localising this centrally as previously proposed [53]. This allows for more detailed tracking of infant body-parts and the associated movements, and subsequently deeper analysis of the correlated motion.

In [53], optical flow data was extracted from video sequences before wavelet frequency analysis was implemented in order to calculate the time-dependant trajectory signals found in the data. However, using this optical flow method presented issues with tracking larger movements, and as such it was suggested that video captured at a higher frame rate would be required for future analysis. In order to address this problem several works [23], [40], [44], [45] implemented large displacement optical flow (LDOF) [8], to track infant body movements and extract features for classification. The LDOF method extends optical flow to better deal with large displacements of foreground objects and camera movement, making it more suitable for detailed infant motion analysis. This enhancement allowed for the improvement of existing motion quantity features [23], the development of velocity-based features [40], [44], and the introduction of frequency-based features [45] for infant motion analysis and binary classification. Through comprehensive analysis of these movement-based features, it was subsequently determined that dynamic features are typically more predictive than the previous statistical features [45]. However, similar challenges are present, such as issues with occlusion, drift and noise [54], as well as susceptibility to unrelated movements (such as equipment, parents or clinicians in shot), and sensitivity to illumination changes [43]. Additionally, the low generalizability and interpretability of extracted features make translation to clinical practice less likely [43], [45]. Finally, the motion data does not easily assess body parts separately, typically relying upon whole body movement analysis [10], [46]. By incorporating part-based assessment, the potential to provide further analytic cues for CP prediction and the identification of CP sub-types is enhanced [43]. As such, alternative methods have been proposed to better model human shape and motion, such as pose estimation [9], [22].

C. Pose Estimation-Based Methods

Pose estimation is the task of using computer vision and machine learning techniques, to detect human figures in images and video, and determine their pose by estimating the spatial locations of key body joints. In this particular use case, it is suggested that pose estimation is more robust to several factors negatively affecting the discussed methods [11], such as illumination changes, camera movement, changes in resolution, inconsistencies in infant size, and larger movements between frames. Since pose-estimation provides localised joint estimation it is also better able to deal with external influences, as well as provide important motion information based upon individual body-part movements, requiring comparatively minimal manual tuning [10], [36].

Research into the feasibility of using pose-based assessment for GMA is ongoing, with several papers contributing advances in this area. In [10], [18], [27], [30] and [36], pose estimation and domain adaptation methods for video sequences of infants are proposed, with each suggesting that their methods are a step towards automated pose-based GMA. However, each typically assesses the effectiveness of the extracted pose rather than the effect this might have upon final classification. In contrast, [33], [34], [38], [51], [56], examine the viability of using pose-based features for the prediction of CP using a feature extraction and classification pipeline.

Whilst reasonable results have been reported in each of the previously discussed related methods, evaluation of the robustness of the proposed features has not been demonstrated across multiple datasets. Indeed, given the sensitive nature of the data required for evaluation of each of the discussed methods, it is difficult to assess their viability without full re-implementation and assessment on shared datasets. As such, we not only propose a wide range of new GMA relevant features capturing orientation, displacement, and frequency information (Section III-E), and a feature fusion pipeline (Section III-G), but also a full re-implementation of several methods for comparative testing across multiple datasets, including videos captured in a real-world clinical setting (Section IV).

III. Methodology

In this section we provide details of the study design, data collection and the composition of the two datasets used for evaluation. We also discuss the proposed framework, data preprocessing, feature extraction, and the classification techniques used.

A. Study Design and Data Collection

For this collaborative project, a retrospective cohort study design was implemented for data collection to produce the

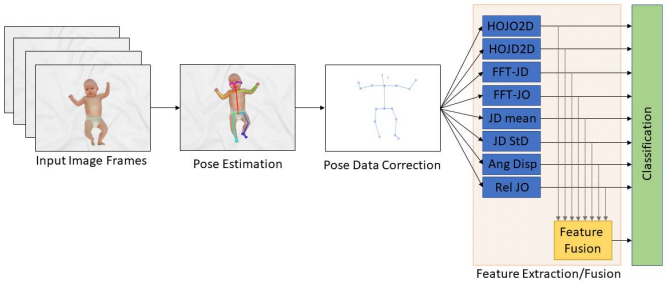


Fig. 1. Overview of the pose estimation (Section III-D), feature extraction (Section III-E) and classification (Section III-F) framework.

RVI-38 dataset (detailed in Section III-B.1). Ethical approval was obtained from the host organisation (Ref: 9865), the Research Ethics Committee (REC), the Health Research Authority (HRA), and Health and Care Research Wales (HCRW) (Ref: 19/LO/0606, IRAS project ID: 252317). Parental consent for the use of video recordings was obtained by the clinical staff associated with the project prior to implementation. The study population included infants who had a clinical GMA, with a video recording at 3–5 months post-term, as part of their routine follow-up care. The MINI-RGBD dataset (detailed in Section III-B.2) is an open source dataset, and all terms specified in the author's license agreement were met.

B. Datasets

In order to comprehensively evaluate our feature extraction, fusion and classification pipeline, we made use of two separate datasets, details of which are discussed in this section. These datasets were also used to evaluate the selected baseline methods for comparison, as discussed in Section IV.

1) RVI-38 Dataset: An important part of this study is that the framework has to have the ability to generalise well across different datasets, particularly when processing real-world video data. To reflect this, the challenging new RVI-38 dataset was collected to inform the design of our framework, and for evaluative analysis. This dataset is composed of real patient video data gathered as part of routine clinical care at the Royal Victoria Infirmary (RVI) in Newcastle upon Tyne, and reflects the genuine intra-class variance and subsequent complexity present in the real-world clinical setting. The RVI-38 dataset consists of 38 videos, of 38 different infants aged between three and five months post-term. The videos were recorded using a handheld Sony DSC-RX100 Advanced Compact Premium Camera with a resolution of 1920×1080 @ 25 FPS. The duration of each video varied between a minimum of 40 seconds and a maximum of 5 minutes, with an average duration of 3 minutes and 36 seconds. The footage was captured from above, in a top-down orientation, with the infant lying in a supine position per GMA guidelines. However, unlike many of the related works (e.g [1], [2], [10], [23], [40]), all video recordings were used as part of our evaluation, with no prior screening for inconsistencies such as poor lighting, camera movement, external factors, significant shadows, or additional selective pre-processing. Whilst this produced a more challenging dataset, it also represents a real-world evaluation of footage captured in a clinical setting. By including this challenging footage we hope to demonstrate that our system is more capable of being robust to variations in data capture, making it more suitable to clinical implementation. The videos were classified by two experienced assessors, using the GMA, into one of two categories; 1) FM+ where the infant demonstrates normal movements indicative of typical development; and 2) FM− where the infant demonstrates abnormal movement patterns that may be of concern to clinicians. This resulted in 32 videos being annotated as FM+ and 6 videos being annotated as FM−.

2) MINI-RGBD Dataset: One of the challenges facing researchers attempting to automate the GMA is the availability of suitable data. Given that the video data required for the GMA is of a sensitive nature, baseline datasets are not currently publicly available. Additionally, since human pose estimation frameworks are almost exclusively trained and tested using images of adults, a dataset consisting of images of infants for research purposes can understandably be difficult to obtain. To address this problem the Moving INfants In RGB-D (MINI-RGBD) dataset was generated and made publicly available [20]. This dataset maps real-world 3D infant movements, captured in a clinical setting, to virtual 3D models of infants. Photo-realistic videos of the 3D infant models were produced using computer graphics rendering, allowing for the generation of anonymised, and subsequently shareable footage, which retains the real-world movement characteristics required for the GMA. This dataset consists of 12 top down videos of infants lying in a supine position, each 40 seconds in duration. The videos were once again analysed by two experienced GM assessors and similarly labelled, resulting in 8 videos being annotated as FM+ and 4 videos being annotated as FM−.

C. Proposed Framework Overview

In this paper, we make use of traditional machine learning classification algorithms for practicality and interpretability (Fig 1). We use raw video as input for the pose estimation framework to compute joint positions, which are then corrected to remove outliers and inconsistencies, as discussed in Section III-D. This corrected data is then used to generate features based upon the GMA for further analysis. Details of the features used and the extraction processes are discussed in Section III-E. These individually extracted features are then fed into a classification framework for evaluation, as discussed in Section III-F. Finally, we fuse the features together to create a more robust representation and evaluate this against both the individual features and the selected baselines, as discussed in Section IV.

D. Pose Estimation and Data Pre-Processing

In order to extract meaningful features for subsequent analysis and classification, pose estimation methods are utilised in this paper. Specifically, we make use of the OpenPose framework [9] to extract joint positions from 2D RGB video data. The extracted joint positions form a skeletal pose representation consisting of 25 predefined joints, as shown in Figure 2. As such, each frame of each video is represented

by 25 sets of 2D (x and y) coordinates and an associated confidence score for the prediction.

In this work, we make use of all the extracted joints with the exception of the facial landmarks (joints 16 to 19), and the feet (joints 20 to 25), as it was determined that these joints were less reliable than the other body landmarks acquired through OpenPose [9] due to self occlusion errors, and were found to play a less important role in the final GMA-based classification results.

1) Automated Data Correction: To ensure consistency throughout the pipeline, the exported OpenPose data is pre-processed prior to feature extraction. This pre-processing involves remapping anomalous joint positions caused by self occlusion or inaccuracies in the OpenPose joint prediction process. The first stage is a qualitative evaluation of the extracted OpenPose data, to check that predicted joint positions and the associated confidence scores correctly align, and are consistent with the input video. We then use the predicted confidence scores to calculate a confidence threshold. The confidence threshold is calculated by taking the average confidence score, per joint, across each video sequence, and subtracting 5% from this. This means that we are able to remove joint positions with a lower confidence score than the confidence threshold, removing outliers on a frame by frame basis, for each joint, in each video sequence. As such, we compute the confidence threshold value t_i for joint i by

$$t_i = (\frac{1}{n}\sum_{j=1}^{n} c_{i,j}) \times 95\% \tag{1}$$

where n is the total number of frames (or postures), $c_{i,j}$ is the confidence score of joint i at frame j returned by OpenPose.

For joints identified as outliers, we calculate a revised joint position based upon the coordinates of the joint in the nearest neighbouring frames with confidence scores higher than the confidence threshold. To calculate the revised joint position we use modified Akima interpolation [4] as proposed in [34].

We then apply a moving-average filter between frames on a per joint basis to reduce jitter present in the sequence. With the filter, smoother, more reliable movements are generated for motion analysis. Empirically, we found that using a filter calculated over a 5 frame sliding window provides the best results.

2) Automated Data Normalisation: To ensure that the orientation and displacement is comparable between videos, we rotate and normalise the landmark coordinates within each frame. Using joint 9 as the root, we amend all joint coordinates so that the root is fixed at 0,0 whilst the relative distance of each of the joints is unchanged. We then calculate the rotation θ^f_{align} required at frame f to align the spinal column (i.e. the central line between joints 2 and 9) with the $y_{axis} = (0, 1)$ by

$$\theta^f_{align} = dir \times \arccos \frac{(fp^f_2 - fp^f_9) \cdot y_{axis}}{\|(fp^f_2 - fp^f_9)\| \|y_{axis}\|} \tag{2}$$

where fp^f_2 and fp^f_9 are the filtered 2D coordinates of joint 2 and 9, respectively, and $dir = sign(fp^f_2 - fp^f_9) \times y_{axis})$ is used to determine the direction of the rotation (i.e. clockwise or counter-clockwise). Finally, the normalized position p of each joint can be computed by

$$p^f_i = \begin{bmatrix} cos(\theta^f_{align}) & -sin(\theta^f_{align}) \\ sin(\theta^f_{align}) & cos(\theta^f_{align}) \end{bmatrix} (fp^f_i - fp^f_9)^T \tag{3}$$

where $i \in [1, 15]$.

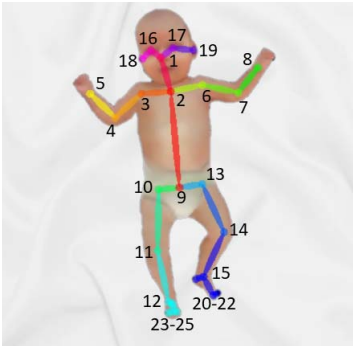


Fig. 2. The 25 joint OpenPose [9] output skeleton, with associated joint reference numbers overlaid on an example input RGB image from the MINI-RGBD dataset [20].

E. Feature Extraction

Motivated by the encouraging results obtained in our pilot study [33], we suggest that pose-based histogram features can effectively represent the motion and distribution of postures over time related to the GMA. Using the corrected and normalised pose data, we therefore propose several new pose-based features for the analysis of infant body movements and subsequent prediction of CP based upon distinct features from the GMA. Given that GM assessors typically look for specific movement patterns, we attempt to model these patterns through a set of orientation-based, displacement-based and frequency-based features. Specifically we aim to model the movements associated with the assessment criteria set out in the GMA checklist [3], and the passive movement assessment section of the Optimality Score neurological examination [19]. In this section we discuss details of the proposed new features, their relevance to the GMA, and the methods used for feature fusion.

1) Angular Displacement (HOAD2D): Angular Displacement represents the change in angular orientation across a specified time interval for each body part in the video. This histogram-based feature captures the distribution of the angular displacement between a predefined regular offset interval. As such, the smoothness of the body part movements can be represented. For example, a smooth movement should be characterised by a histogram which has only a small number of bins with high values. This feature is therefore designed to help identify spasmodic, abrupt, and sporadic movements of short duration. The orientation of each joint is the 2D vector pointing from the parent joint to the child joint:

$$o^f_i = p^f_i - p^f_j \tag{4}$$

where o^f_i is the orientation (2D vector) of joint i at frame f, p^f_i and p^f_j are the 2D coordinates of joint i and j, respectively,

with joint j being the parent of joint i. This method of joint orientation calculation was also used to extract the HOJO2D feature in our pilot study [33], however, here, we further compute the angular displacement θ_i^f of joint i at frame f by:

$$\theta_i^f = \arccos \frac{o_i^f \cdot o_i^{f-\Delta t}}{\|o_i^f\| \|o_i^{f-\Delta t}\|} \tag{5}$$

where Δt is a predefined regular offset interval of 10 frames, the value of which we determined empirically from our experiments. Equation 5 represents the cosine similarity of o_i^f and $o_i^{f-\Delta t}$. By applying this method, the direction (i.e. clockwise or counter-clockwise) of the angular displacement is discarded such that the feature solely focuses on the magnitude of the orientation change.

Having computed the angular displacements of every joint in the whole pose sequence, the HOAD2D for each joint is computed by quantizing the displacements into a finite number of bins. Therefore, the number of bins and the size (i.e. range of values) of each bin will significantly affect the discriminative power of the feature. We observed that most of the angular displacements are very small, while the maximum theoretical displacement is 180°. As a result, we propose using non-uniform bin sizes to better represent the distribution for the angular displacements:

$$bs_i = \frac{180^\circ}{2^{n-i}} \tag{6}$$

where bs_i is the bin size for the i-th bin and n is the total number of bins for the histogram feature. For HOAD2D, we empirically found that $n = 16$ yields the best results.

2) Relative Joint Orientation (HORJO2D): To analyse the coordination and synchronisation of different body part movements, it is important to extract features from different joints simultaneously. Inspired by [50], we propose representing the distribution of the relative orientation of the joints using a histogram-based feature. Here, the pairwise relative joint orientation is computed in a similar manner as in Equation 4:

$$o_{i\to j}^f = p_j^f - p_i^f, \tag{7}$$

although the two joints are not necessarily physically connected. In order to capture the synchronisation of different parts of the body, we compute the relative orientation for all pairs of joints.

Since the relative joint orientation has a potential range of 0° to 359°, a uniform bin size is used. In doing so, we empirically found that $n = 16$ produced the best performance. Once the individual joint histograms have been extracted, we combine these to form histogram representations for each limb prior to concatenation for classification. HORJO2D intuitively represents the body synchronisation, as such, a histogram which has only a small number of bins with high values means that the joints are moving in the same direction together.

3) Relative Joint Angular Displacement (HORJAD2D): To further capture the change in body part movement synchronisation over time, the angular displacement of the relative joint orientation is also extracted as a histogram feature. This feature is crafted to evaluate the relationship between body parts, such that whole body coordination, dystonia, and ataxic movements can be assessed. As with extracting the HORJO2D feature, the pairwise relative joint orientation (RJO) is computed and similarly combined. We further compare the RJOs before and after the predefined frame offset interval, and angular displacement is calculated using the cosine similarity of the two RJO vectors similar to the calculation of HOAD2D, as in Equation 5.

Again, most of the angular displacements computed have a small values. As a result, we again implement a non-uniform bin size (Equation 6) to increase the discriminative power of the HORJAD2D feature. From our experiments, we empirically found that the best results were obtained when $n = 8$.

4) Fast Fourier Transform of Joint Displacement (FFT-JD): Whilst the aforementioned histogram features represent the distribution of different kinds of spatial features at a coarse level, information pertaining to temporal ordering is discarded. Inspired by previous work analysing body movements in the frequency domain [45], we propose the FFT-JD feature. This feature contains the magnitude of each of the frequency components extracted from the motion such that the variability of the motion can be better assessed. By using the Fast Fourier Transform (FFT) we convert the extracted joint displacement signal $D_i = [\|\dot{p}_i^2\|, \|\dot{p}_i^3\|, \cdots, \|\dot{p}_i^m\|]$ of joint i from a motion with m frames to a representation in the frequency domain, allowing us to model the complexity, fluidity, and variety of the movements, whilst highlighting any repetitive, athetoid, tremulous, or myoclonic characteristics. Additionally it is reported that analyzing human motion in the frequency domain is more robust to noisy data [55], and as such helps with the task of assessing some of the smaller, more detailed movements associated with the GMA.

We extract the FFT-JD by applying FFT to the vector D_i:

$$Y_i^k = \left| \frac{1}{m} \sum_{f=0}^{m-1} D_i e^{\frac{-l2\pi kf}{m}} \right| \tag{8}$$

where Y_i^k is a vector which contains the magnitude of the frequency component at index k for joint i, $e^{\frac{l2\pi}{m}}$ is a primitive m^{th} root of 1.

Having computed the frequency component Y_i from D_i, Y_i is partitioned into 16 bins with non-uniform bin sizes:

$$bs_{FFT-JD,b} = \begin{cases} F\frac{b^2}{n^2}, & \text{if } b = 1, \\ F\frac{b^2}{n^2} - \sum_{k=1}^{b-1} bs_{FFT-JD,k}, & \text{if } 2 \le b < n, \\ F - \sum_{k=1}^{b-1} bs_{FFT-JD,k}, & \text{if } b = n. \end{cases} \tag{9}$$

where $bs_{FFT-JD,b}$ is the size of the b-th bin and F is the number of frequency components obtained from D_i using FFT. The last bin (i.e. $bs_{FFT-JD,n}$) will occupy the remaining space.

5) Fast Fourier Transform of Joint Orientation (FFT-JO): Similar to the FFT-JD feature we once again make use of FFT to model repetitive movements by looking into the frequency components. This feature provides information relating to the rigidity, directional variation, and range of movement associated with the infant's posture. In this case we model the

repetition and frequency of similar postures from a joint orientation sequence $O_i = [o_i^1, o_i^2, \cdots, o_i^m]$ for joint i using FFT as in Equation 8. The histogram-based FFT-JO is computed using the same method described in Equation 9, where the frequency components are computed using O_i instead. In this case, the bins for the lower frequency components will be smaller and can more effectively represent the more relevant low frequency components, as opposed to the high frequency motion signals which potentially contain noise.

6) Histograms of Joint Orientation (HOJO2D) and Histograms of Joint Displacement (HOJD2D): In this work we improve upon our previously reported method [33], by extracting individual HOJO2D and HOJD2D joint histograms and concatenating these to form limb-based representations. This method means that we are able to incorporate a greater range of motion detail than the previous method, which extracts an individual per-limb histogram representation grouping several joints together.

F. Classification

Once we have extracted features, we then used the Z-score to standardise the feature data h, to ensure it is on the same scale prior to further analysis by $z = \frac{h-\mu}{\sigma}$, where μ and σ are the mean and standard deviation of all samples in the training set, and z contains the normalized features. In our implementation we use the Z-score as this allows our system to retain the shape properties of the original data set, with our initial classification results showing improvements using this method over min-max normalization. We then feed the features into a classification framework to obtain an overall prediction on the prevalence of CP based upon the annotations provided by the GMA assessors (Section III-B). Our classification framework consists of several machine learning algorithms, namely Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (Tree), Linear Discriminant Analysis (LDA), Ensemble of classification models (Ens), and k-Nearest Neighbour where $k = 1$ and $k = 3$. This approach allows us to generate a suitable interpretation of the strength of the features assessed, and the performance of each classifier in this task. The metrics used to evaluate the features and associated classification performance are discussed in Section IV-B.

G. Feature Fusion

In addition to evaluating the classification performance of each of the individual features, we also fuse our selected features together for further analysis (see Fig. 1). This fusion process concatenates two separate feature sets into *pose-based* features, and *velocity-based* features which are used for classification. The pose-based features represent the angular feature information extracted from the pose data, as such these representations are indicative of the overall quality of the infant posture and the predominant directions of movement. The pose-based features consist of a concatenation of HOJO2D, HOAD2D, HORJO2D, and FFT-JO. The velocity-based features represent the displacement of the joints over predefined time intervals, and as such model the speed, fluidity, coordination, and complexity of the infant movements. The velocity-based features consist of a concatenation of HOJD2D, HORJAD2D, and FFT-JD. Lastly, we fuse the pose-based feature set with the velocity-based feature set for classification. By concatenating the features through early fusion, it is expected that improved classification performance will be achieved, provided the classifier is capable of handling the higher dimensionality input data.

IV. Evaluation

In this section we provide details of our evaluation methods. In Section IV-A we discuss the baseline methods used for comparison with our proposed method. In Section IV-B we discuss the metrics used for our comparative evaluation. In Section IV-C we discuss the experimental settings used and the rationale behind each selected test. In Section IV-D we discuss our classification results and in Section IV-E we examine the hyperparameter optimisation. Finally, in Section IV-F we include an analysis of the proposed features.

A. Baseline Methods

In order to assess the effectiveness and robustness of our system, we reimplement several video-based methods from the literature to serve as baselines for comparison, including Centroid of Motion and Quantity of Motion [1], [2]; Absolute Motion Distance, Relative Frequency, and Magnitude of Wavelet Coefficients [53]; and Frequency Analysis [45]. We also compare our results with those reported in [56] and [57], as well as conducting our experiments using the source code provided by the authors of [38] and [51].

1) Centroid of Motion: The centroid of motion is the spatial centre point of the motion image which highlights the pixels with detected changes (i.e. body movement), and in our case, represents the centre point of the movements of the infant. As discussed in [2], the mean and standard deviation of centroid of motion in the x-and y-directions (CX_{m}, CX_{SD}, CY_{m} and CY_{SD}) are calculated and exported as features for classification.

2) Quantity of Motion: The quantity of motion is also calculated through the generation of a motion image [2]. It is the sum of all pixels with positive values from the motion image, divided by the total number of pixels contained within the image. The standard deviation ($\mathrm{Q_{SD}}$) and mean ($\mathrm{Q_m}$) of the quantity of motion are therefore calculated and used for classification.

3) Cerebral Palsy Predictor: We also include the Cerebral Palsy Predictor (CPP) feature set as discussed in [2]. This is the concatenated combination of the centroid of motion standard deviation ($\mathrm{C_{SD}}$), the quantity of motion mean ($\mathrm{Q_m}$), and the quantity of motion standard deviation ($\mathrm{Q_{SD}}$).

4) Absolute Motion Distance: The Absolute Motion Distance (AMD), Relative Frequency (RF), and Magnitude of Wavelet Coefficients (MWC) methods are all based upon optical flow information, which is used for motion-based tracking. We followed the technical details presented in [53] in our implementation. The AMD is proposed as a holistic measure of activity, as it captures the absolute values of the optical flow velocities and stores them in histogram format. Since the bin size is not specified in [53], we experiment with bin sizes

of 8, 16, 32, 64 and 128 in our implementation. We reported the best results obtained using the bin size of 8 in this paper.

5) Relative Frequency: The relative frequency (RF) of the signal represents the occurring frequencies found in the movement patterns, which are converted into a histogram representation for classification. In our implementation, we followed [53] and again used bin sizes of 8, 16, 32, 64 and 128 since this is not specified. From our experiments we empirically found a bin size of 8 returns the best results.

6) Magnitude of Wavelet Coefficients: The Magnitude of Wavelet Coefficients (MWC) power spectrum is used to demonstrate the variety of the observed movement at different resolution levels, providing insight into the complexity of the movement. We followed [53] to compute the associated histogram features.

7) Frequency Analysis: Given that normal FMs are defined as an ongoing and variable stream of movements, [45] suggest that these motions can be better studied in frequency domain. As such, the Fast Fourier Transform (FFT) was used to obtain the frequency components of the motion. We followed [45] to extract the mean and standard deviation values of the Fourier coefficients in horizontal ($FFTx_m$ and $FFTx_{SD}$) and vertical directions ($FFTy_m$ and $FFTy_{SD}$) as features for classification, using 100 bins with non-uniform sizes as specified in the literature.

B. Evaluation Metrics

In this paper we make use of several evaluation metrics to assess the performance of each feature and the associated classifier. In our evaluation, true positive (TP) is a measure of the cases in which impaired infants are correctly classified as impaired, true negative (TN) represents unimpaired infants correctly classified as unimpaired, false positive (FP) represents unimpaired infants incorrectly classified as impaired, and false negative (FN) represents impaired infants incorrectly classified as unimpaired. Based upon these metrics, the sensitivity (SE) is defined as the percentage of correctly identified positive classifications amongst the positive population of the dataset, the specificity (SP) is the percentage of correctly identified negative classifications amongst the negative population of the dataset, and the accuracy (AC) is defined as the percentage of all correctly classified instances. We also calculate the precision (PR), recall (RE), and F1 Score (F1). PR represents the percentage of correctly identified positive cases from all positive predictions, RE measures the correctly identified positive cases from all of the actual positive cases, and F1 is the harmonic mean of PR and RE and typically provides one of the most valuable measures of performance [13]. In addition to this, we also calculate the Matthews Correlation Coefficient (MCC) [32], which is regarded as a reliable measure where there are significant differences between the classes sizes, and as such it is widely used as a consistent performance metric on imbalanced datasets [7], [12].

C. Experimental Settings

To evaluate the generality of the classifiers using both our proposed features and the baseline approaches, a leave-one-subject-out cross validation method is used. Splitting the data in this manner ensures that the classifiers will be evaluated using unseen data, as suggested in previous works [33], [34], [56]. We further evaluate the performance of each type of feature on different classifiers (as explained in Section III-F). We report the best result for each method along with the associated classifier.

TABLE I
CLASSIFICATION RESULTS USING THE MINI-RGBD DATASET [20]

Feature	Class.	AC	SE	SP	F1	MCC
CX_m [2]	Ens	83.33	50.00	**100.00**	66.67	63.25
CX_{SD} [2]	Ens	83.33	75.00	87.50	75.00	62.50
CY_m [2]	LDA	33.33	**100.00**	0.00	50.00	0.00
CY_{SD} [2]	k=3	75.00	75.00	75.00	66.67	47.81
Q_m [2]	Ens	58.33	50.00	62.50	44.44	11.95
Q_{SD} [2]	k=3	75.00	75.00	75.00	66.67	47.81
CPP [1]	Tree	66.67	75.00	62.50	60.00	35.36
AMD [53]	LDA	91.67	**100.00**	87.50	88.89	83.67
MWC [53]	LDA	83.33	75.00	87.50	75.00	62.50
RF [53]	LDA	91.67	**100.00**	87.50	88.89	83.67
$FFTx_m$ [45]	Tree	83.33	75.00	87.50	75.00	62.50
$FFTx_{SD}$ [45]	Ens	58.33	50.00	62.50	44.44	11.95
$FFTy_m$ [45]	Tree	91.67	75.00	**100.00**	85.71	81.65
$FFTy_{SD}$ [45]	Tree	75.00	75.00	75.00	66.67	47.81
FFT_m [45]	Tree	83.33	75.00	87.50	75.00	62.50
FFT_{SD} [45]	LR	75.00	75.00	62.50	72.73	59.76
MCI [56]	n/a	91.67	**100.00**	87.50	88.89	83.67
CA [57]	DNN	91.67	-	-	-	-
BPB [51]	DNN	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
STAM [38]	DNN	91.67	**100.00**	87.50	88.89	83.67
HOJO2D	Ens	91.67	75.00	**100.00**	85.71	81.65
HOJD2D	Ens	83.33	75.00	87.50	75.00	62.50
FFT-JO	Ens	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
FFT-JD	LR	91.67	75.00	**100.00**	85.71	81.65
HOAD2D	LR	66.67	**100.00**	50.00	66.67	50.00
HORJO2D	LDA	91.67	75.00	**100.00**	85.71	81.65
HORJAD2D	Ens	83.33	75.00	87.50	75.00	62.50
Pose	Ens	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Velocity	Ens	91.67	**100.00**	87.50	88.89	84.32
Pose & Vel.	Ens	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

D. Experimental Results

The classification results on the MINI-RGBD and RVI-38 datasets are presented in Table I and Table II, respectively. From our evaluation, we observe that our individual features are typically performing at a similar level to that of the best features proposed in the related works. However, we note that the fusion of our proposed features is providing state-of-the-art performance across both of the datasets used in our experiments, as discussed the following subsections.

1) Classification Performance on Individual Features: From Table I, it can be seen that our proposed feature FFT-JO achieved a 100.00% classification accuracy on the MINI-RGBD dataset. Only one of the 20 baseline methods (BPB) evaluated achieved this perfect classification result in our tests, highlighting the remarkable performance of this new feature. Encouraging results are also obtained using our other frequency-based feature FFT-JD with 91.67% classification accuracy, 85.71% F1 score, and 81.65% MCC. This performance is higher than all of the 20 baselines in the experiments, with the exception of AMD (F1:88.89%, MCC:83.67%), RF (F1:88.89%, MCC:83.67%), FFT-Y_m (F1:85.71%, MCC:81.65%), MCI (F1:88.89%, MCC:83.67%),

and STAM (F1:88.89%, MCC:83.67%). Similarly, our newly proposed HORJO2D feature achieved the same performance of 91.67% classification accuracy, 85.71% F1 score, and 81.65% MCC. However, the other relative orientation-based feature HORJAD2D is not performing as well on this dataset with 83.33% classification accuracy, an F1 Score of 75% and MCC of 62.50%. Although this performance still outperforms most of the baselines, the noticeably lower specificity (87.50%) results in a lower overall classification performance for this feature. For the angular displacement based feature HOAD2D, an average performance is obtained on this dataset with an F1 score of 66.67%, matching or outperforming 8 of 20 baselines.

For the RVI-38 dataset, we note a general drop in performance due to the challenging nature of the dataset, as shown in Table II. This is particularly noticeable in the baseline methods where we see a significant drop for each baseline, with the exception of MWC (F1:75%, MCC:62.50%). This drop is most likely associated with the challenging nature of the captured data and the full frame analysis of these methods. We are seeing that methods which are able to deal with external influences better, such as the pose-based methods, are generally producing more accurate results. This is also reflected in the results produced using our proposed individual features. In this setting we note that the HOAD2D feature is performing particularly well, representing the strongest individual feature on this dataset, recording the highest F1 Score (83.33%) and MCC (80.21), along with the joint highest accuracy (94.74%) and sensitivity (83.33%). The reworked HOJO2D (F1:72.73%, MCC:68.54) and HOJD2D (F1:80.00%, MCC:79.21%) again perform well, showing the robustness of these improved features. The HORJO2D feature is also performing well, with an accuracy of 92.11%, F1 score of 76.92%, and MCC of 72.51%. We note that FFT-JD is once again performing well, with an accuracy of 92.11%, an F1 score of 72.73%, and MCC of 68.54%. We also observe that whilst FFT-JO and HORJA2D achieve a reasonable performance on the RVI-38 dataset, with 84.21% and 86.84% classification accuracy respectively, the F1 and MCC scores are lower than our other proposed features. However, whilst the scores for these features are not class leading, they are still higher than those achieved by 16 of the 18 baseline methods evaluated in this setting.

2) Classification Performance on Feature Fusion: We observe that on the MINI-RGB dataset the pose-based fusion is extracting the strongest feature representation and retaining the perfect classification performance provided by the FFT-JO individual feature. Our evaluation also suggests that whilst the pose-based fused features are generally outperforming the velocity-based features, fusing both of these feature sets further improves performance on both the MINI-RGB dataset (F1: 100%, MCC 100.00%) and the RVI-38 dataset (F1: 90.91%, MCC: 89.89%). We note that on the RVI-38 dataset, the strengths from each feature set combine to provide this improved overall classification performance, with the higher sensitivity found in the pose-based features (83.33%) and the higher specificity found in the velocity-based features (100.00%) directly translating to the concatenated fusion of these feature sets. This observation aligns well with our feature design, which looks to incorporate the combined positional, directional, postural, and transitory information specified in the GMA guidelines. We also note that on the fused features we are seeing a consistently high performance using the Ensemble classifier, with the best results obtained on both datasets for all fused feature sets using this classification method. The evaluation metrics also highlight the robustness of the proposed feature fusion method, given that only one positive sample video was misclassified across both datasets.

TABLE II
CLASSIFICATION RESULTS USING THE RVI-38 DATASET

Feature	Class.	AC	SE	SP	F1	MCC
CX_m [2]	LR	50.00	**83.33**	43.75	34.48	20.20
CX_{SD} [2]	Ens	68.42	33.33	75.00	25.00	6.90
CY_m [2]	k=3	84.21	50.00	90.63	50.00	40.63
CY_{SD} [2]	LR	63.16	66.67	65.63	36.36	21.54
Q_m [2]	LR	52.63	50.00	53.13	25.00	2.28
Q_{SD} [2]	k=1	86.84	50.00	93.75	54.44	47.19
CPP [1]	Ens	84.21	50.00	90.63	50.00	40.63
AMD [53]	LDA	83.33	50.00	**100.00**	66.67	63.25
MWC [53]	Tree	83.33	75.00	87.50	75.00	62.50
RF [53]	LDA	84.21	66.67	87.50	57.14	48.45
FFT-X_m [45]	Ens	84.21	50.00	90.63	50.00	40.63
FFT-X_{SD} [45]	LR	63.16	66.67	62.50	36.36	21.54
FFT-Y_m [45]	k=1	81.58	33.33	90.63	36.36	25.84
FFT-Y_{SD} [45]	LDA	55.26	50.00	56.25	26.09	4.58
FFT_m [45]	Tree	84.21	50.00	90.63	50.00	40.63
FFT_{SD} [45]	LR	42.11	66.67	37.50	26.67	3.15
BPB [51]	DNN	84.21	33.33	93.75	40.00	32.18
STAM [38]	DNN	81.58	33.33	90.63	36.36	25.85
HOJO2D	Ens	92.11	66.67	96.88	72.73	68.54
HOJD2D	Ens	94.74	66.67	**100.00**	80.00	79.21
FFT-JO	LDA	84.21	**83.33**	84.38	62.50	56.07
FFT-JD	Ens	92.11	66.67	96.88	72.73	68.54
HOAD2D	Ens	94.74	**83.33**	96.88	83.33	80.21
HORJO2D	Tree	92.11	**83.33**	93.75	76.92	72.51
HORJAD2D	LR	86.84	66.67	90.63	61.54	53.89
Pose	Ens	94.74	**83.33**	96.88	83.33	80.21
Velocity	Ens	94.74	66.67	**100.00**	80.00	79.21
Pose & Vel.	Ens	**97.37**	**83.33**	**100.00**	**90.91**	**89.89**

E. Hyperparameter Optimisation

To refine the framework performance, we further investigate hyper-parameter optimisation using Bayesian Optimisation [52]. Using this method, we evaluate the results of optimisation in an informed manner, by tuning the learning rate, the number of learning cycles, the minimum observations per leaf, and the maximum number of branch nodes, to minimize the cross-validation loss of the classifier.

We present several plot representations of our Bayesian Optimisation in the supplementary information. In the plots, the number of function evaluations relates to the iteration number of the objective function, the min objective is the minimum value that the objective function has reached up to the current iteration, and the estimated minimum objectives are the mean values of the posterior distribution of the Gaussian process model of the objective function [25]. We also map the hyper-parameter variables to the classification performance metrics to determine the optimal hyper-parameters. In this setting, we found the optimal hyper-parameters to be: 0.1045800 learning rate, 11 learning cycles, 1 minimum observation per leaf, and 32 split branch nodes, providing an objective function of 0.026316 and an accuracy of 100% on

TABLE III
THE P-VALUES OF THE FEATURES COMPUTED FROM CHI-SQUARE TESTS ON THE MINI-RGBD AND RVI-38 DATASETS

		MINI-RGBD			RVI-38		
Feature	**Dim.**	**Median**	**# *sp***	**% *sp***	**Median**	**# *sp***	**% *sp***
AMD [53]	8	0.1020	0	0%	0.1280	2	25%
MWC [53]	18	0.6456	0	0%	0.8931	6	33.33%
RF [53]	16	0.1020	4	25.00%	0.2388	6	37.50%
HOJO2D	64	0.1020	18	28.13%	0.1370	10	31.25%
HOJD2D	128	0.0860	52	**40.63%**	0.1849	45	35.16%
FFT-JO	48	0.1020	8	16.67%	0.1849	18	32.14%
FFT-JD	64	0.1020	23	35.94%	0.1849	18	28.13%
HOAD2D	64	**0.0847**	17	26.56%	0.1849	22	34.38%
HORJO2D	32	0.1020	5	15.63%	**0.050**	14	**43.75%**
HORJAD2D	32	0.1020	9	28.13%	0.071	8	25%
Pose & Vel.	792	0.1020	228	28.78%	0.1849	260	32.83%

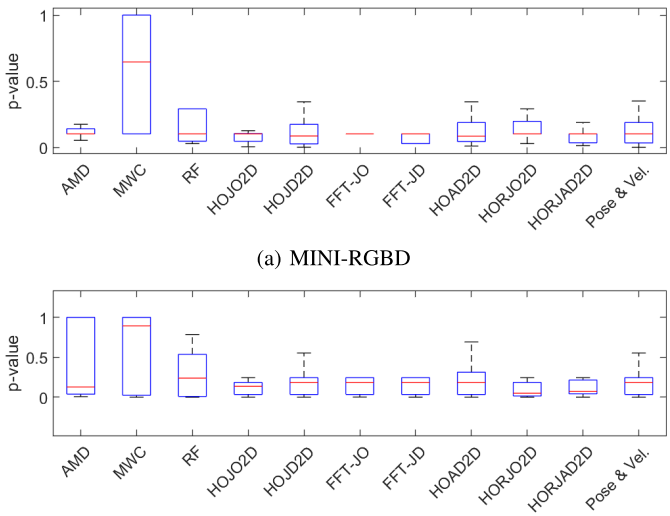
the MINI-RGBD dataset, and 97.37% on the RVI-38 dataset, per our reported results.

F. Feature Analysis

To further evaluate the discriminative power of the newly proposed features, chi-square tests are used for testing if the predictor variables (i.e. the multi-dimensional features proposed in this work) and the response variable (i.e. the label of each video) are related. In particular, such tests have been widely used for feature selection and are thus able to reflect the quality of the features we propose. We further conducted the chi-square tests on both the MINI-RGBD and the RVI-38 datasets to highlight the differences between these two datasets. Specifically, the p-value for each predictor variable is calculated and the median values are reported in Table III. Here, we consider the predictor variables as *significant predictor (sp)* if $p < 0.05$. Since the features used in the experiments are mostly multi-dimensional, the dimensionality (dim.), number of *sp* (# *sp*) and percentage of *sp* (% *sp*) are also reported in Table III. We also include the top performing baselines from our experiments (i.e. AMD, MWC and RF proposed in [53]) in this analysis to further highlight the effectiveness of the proposed features.

From Table III, it can be seen that HOJD2D and HORJO2D achieve the highest % *sp* on MINI-RGBD and RVI-38, respectively. On MINI-RGBD, our proposed features are having the same or lower median p-value when compared with AMD, MWC and RF. On RVI-38, HORJO2D and HORJAD2D achieved significantly lower median p-values than other top performers. Also, it can be seen that AMD performed better than some of our proposed features in terms of the median p-values. However, the results from the 2 datasets also indicate the robustness of our proposed features since more consistent results are obtained by using our features.

To better visualize the distribution of the p-values, boxplots of the p-values of different features on MINI-RGBD and RVI-38 are shown in Figure 3a and 3b, respectively. In particular, the maximum, minimum, first quartile, third quartile and median (red line) values are illustrated in the figures. It can be seen that the majority (from the first to third quartile) of the predictor variables in our proposed features are having a small range with low p-values. This indicates the majority of our proposed features are of higher importance and quality when compared with the AMD, MWC and RF.



(a) MINI-RGBD

(b) RVI-38

Fig. 3. Boxplots of the p-values of different features on each dataset.

V. DISCUSSION

As discussed in Section II-C, we suggest that pose-estimation based approaches provide several advantages over the previously proposed methods in data acquisition and analysis. Our pose and velocity-based method is simpler to understand, retains understandable information, and has less parameters to tune than the related methods, making it more accessible in a clinical setting. We also suggest that, due to the relative assessment of joint motion, our framework is better able to deal with camera movement, changes in resolution, variable infant sizes, and larger motion changes between frames.

However, whilst the results we have achieved in this paper are encouraging, there are several key areas which we would like to address in future works. Our method is heavily dependant upon the quality of the extracted joint positions and as such the pose estimation method. In this paper we made use of OpenPose [9], however this framework is trained using adult data and as such the extracted pose has the potential to have some inconsistencies. Although the results demonstrated here suggest that these inconsistencies are largely dealt with by our qualitative assessment and automated pre-processing techniques, we would like to enhance the pose-estimation by integrating domain adaptation to make the framework more specific to infant body dimensions and posture. We would also like to further explore the temporal aspect of the GMA by extracting spatio-temporal features for further analysis.

Our results on the RVI-38 dataset in particular represent a particularly robust performance given the difficulty of the associated dataset, with only one misclassified video. In this case the misclassified video was one of the positive samples which, in practice presents a greater issue than a misclassified negative sample. In future works we will look to methods

by which we can improve the sensitivity of the classification performance, perhaps through data augmentation to help deal with the moderate class imbalance found in the datasets.

VI. Conclusion

In this paper we have proposed and evaluated several new interpretable motion features relating directly to the criteria associated with the GMA checklist [3] and optimality score [19]. Additionally, we have fused these features together to produce a more robust representation of infant body movement for classification. We compared these features with several other methods from the literature by re-implementing them for assessment using shared datasets. The datasets used in the study consist of the publicly available MINI-RGBD dataset [20], and the proposed RVI-38 dataset, a challenging new video dataset gathered as part of routine clinical care. We find that our proposed fused features achieve state-of-the-art performance across both datasets whilst retaining clinical interpretability. Additionally, we also suggest that by utilising pose-based features, we make the likelihood of collaborative working within the healthcare domain more viable, due to the inherently anonymised and unidentifiable patient data. As such, we make the pose data, and our feature extraction and classification code available to the community to further improve related research.

In future works we hope to explore improving the interpretability of CP prediction models by allowing clinicians to form part of the feedback loop. We will also explore methods of extracting more detailed information about the movement characteristics, through temporal analysis and improved annotation.

Acknowledgment

The authors would like to thank the extended team at the RVI, and in particular, Patricia Dulson for her work annotating and transferring the video data for analysis.

References

[1] L. Adde, J. L. Helbostad, A. R. Jensenius, G. Taraldsen, and R. Støen, "Using computer-based video analysis in the study of fidgety movements," *Early Hum. Develop.*, vol. 85, no. 9, pp. 541–547, Sep. 2009.

[2] L. Adde *et al.*, "Characteristics of general movements in preterm infants assessed by computer-based video analysis," *Physiotherapy Theory Pract.*, vol. 34, no. 4, pp. 286–292, Apr. 2018.

[3] C. Y. P. Aizawa, C. Einspieler, F. F. Genovesi, S. M. Ibidi, and R. H. Hasue, "The general movement checklist: A guide to the assessment of general movements during preterm and term age," *J. Pediatria*, vol. 97, no. 4, pp. 445–452, Jul. 2021.

[4] H. Akima, "A new method of interpolation and smooth curve fitting based on local procedures," *J. ACM*, vol. 17, no. 4, pp. 589–602, Oct. 1970.

[5] M. Bax *et al.*, "Proposed definition and classification of cerebral palsy, April 2005," *Develop. Med. Child Neurol.*, vol. 47, no. 8, pp. 571–576, 2005.

[6] M. Bosanquet, L. Copeland, R. Ware, and R. Boyd, "A systematic review of tests to predict cerebral palsy in young children," *Develop. Med. Child Neurol.*, vol. 55, no. 5, pp. 418–426, May 2013.

[7] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," *PLoS ONE*, vol. 12, no. 6, pp. 1–17, 2017.

[8] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 41–48.

[9] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[10] C. Chambers *et al.*, "Computer vision to automatically assess infant neuromotor risk," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2431–2442, Nov. 2020.

[11] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897.

[12] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[13] R. Cunningham, M. B. Sánchez, P. B. Butler, M. J. Southgate, and I. D. Loram, "Fully automated image-based estimation of postural point-features in children with cerebral palsy using deep learning," *Roy. Soc. Open Sci.*, vol. 6, no. 11, Nov. 2019, Art. no. 191011.

[14] E. S. Draper *et al.*, *MBRRACE-U.K.—Perinatal Mortality Surveillance Report 2017*, Infant Mortality Morbidity Stud., Dept. Health Sci., Univ. Leicester, U.K., 2020.

[15] C. Einspieler and H. F. R. Prechtl, *Prechtls Method on the Qualitative Assessment of General Movements in Preterm, Term, and Young Infants*. London, U.K.: EnglandMac Keith Press, 2004.

[16] C. Einspieler, F. R. H. Prechtl, F. Ferrari, G. Cioni, and F. A. Bos, "The qualitative assessment of general movements in preterm, term and young infants," *Early Hum. Develop.*, vol. 50, no. 1, pp. 47–60, 1997.

[17] *NICE Seeks to Improve Diagnosis and Treatment of Cerebral Palsy*, National Institute for Health and Care Excellence, London, U.K., Jan. 2017.

[18] D. Groos, H. Ramampiaro, and E. A. Ihlen, "EfficientPose: Scalable single-person pose estimation," *Int. J. Speech Technol.*, vol. 51, no. 4, pp. 2518–2533, Apr. 2021.

[19] L. Haataja *et al.*, "Optimality score for the neurologic examination of the infant at 12 and 18 months of age," *J. Pediatrics*, vol. 135, no. 2, pp. 153–161, Aug. 1999.

[20] N. Hesse, C. Bodensteiner, M. Arens, G. U. Hofmann, R. Weinberger, and A. S. Schroeder, "Computer vision for medical infant motion analysis: State of the art and RGB-D data set," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Sep. 2018, pp. 32–49.

[21] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*.

[22] Y. Huang, P. H. H. Shum, S. L. E. Ho, and N. Aslam, "High-speed multi-person pose estimation with deep feature transfer," *Comput. Vis. Image Understand.*, vol. 197, Aug. 2020, Art. no. 103010.

[23] E. A. F. Ihlen *et al.*, "Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study," *J. Clin. Med.*, vol. 9, no. 1, p. 5, Dec. 2019.

[24] F. Y. Ismail, A. Fatemi, and M. V. Johnston, "Cerebral plasticity: Windows of opportunity in the developing brain," *Eur. J. Paediatric Neurol.*, vol. 21, no. 1, pp. 23–48, Jan. 2017.

[25] A. Kazikova, M. Pluhacek, and R. Senkerik, "How does the number of objective function evaluations impact our understanding of metaheuristics behavior?" *IEEE Access*, vol. 9, pp. 44032–44048, 2021.

[26] A. K. L. Kwong, T. L. Fitzgerald, L. W. Doyle, J. L. Y. Cheong, and A. J. Spittle, "Predictive validity of spontaneous early infant movement for later cerebral palsy: A systematic review," *Develop. Med. Child Neurol.*, vol. 60, no. 5, pp. 480–489, May 2018.

[27] M. Li, F. Wei, Y. Li, S. Zhang, and G. Xu, "Three-dimensional pose estimation of infants lying supine using data from a Kinect sensor with low training cost," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6904–6913, Mar. 2021.

[28] H. Mactier *et al.*, "Perinatal management of extreme preterm birth before 27 weeks of gestation: A framework for practice," *Arch. Disease Childhood Fetal Neonatal Ed.*, vol. 105, no. 3, pp. 232–239, May 2020.

[29] N. Maitre, "Skepticism, cerebral palsy, and the general movements assessment," *Develop. Med. Child Neurol.*, vol. 60, no. 5, p. 438, May 2018.

[30] V. Marchi *et al.*, "Automated pose estimation captures key aspects of general movements at eight to 17 weeks from conventional videos," *Acta Paediatrica Int. J. Paediatrica*, vol. 108, no. 10, pp. 1817–1824, 2019.

[31] C. Marcroft, A. Khan, N. D. Embleton, M. Trenell, and T. Plötz, "Movement recognition technology as a method of assessing spontaneous general movements in high risk infants," *Frontiers Neurol.*, vol. 5, p. 284, Jan. 2015.

[32] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[33] K. D. McCay, E. S. L. Ho, C. Marcroft, and N. D. Embleton, "Establishing pose based features using histograms for the detection of abnormal infant movements," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5469–5472.

[34] 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. 51582–51592, 2020.

[35] K. D. McCay *et al.*, "Towards explainable abnormal infant movements identification: A body-part based prediction and visualisation framework," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.

[36] S. Moccia, L. Migliorelli, V. Carnielli, and E. Frontoni, "Preterm infants' pose estimation with spatio-temporal features," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 8, pp. 2370–2380, Aug. 2020.

[37] C. Morgan, I. Novak, C. R. Dale, A. Guzzetta, and N. Badawi, "Single blind randomised controlled trial of GAME (Goals-Activity-Motor Enrichment) in infants at high risk of cerebral palsy," *Res. Develop. Disabilities*, vol. 55, pp. 256–267, Aug. 2016.

[38] 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 Informat.*, vol. 25, no. 10, pp. 3911–3920, Oct. 2021.

[39] J. Olsen, P. Marschik, and A. Spittle, "Do fidgety general movements predict cerebral palsy and cognitive outcome in clinical follow-up of very preterm infants?" *Acta Paediatrica*, vol. 107, no. 2, pp. 361–362, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/apa.14126, doi: 10.1111/apa.14126.

[40] S. Orlandi *et al.*, "Detection of atypical and typical infant movements using computer-based video analysis," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 3598–3601.

[41] M. Oskoui, F. Coutinho, J. Dykeman, N. Jetté, and T. Pringsheim, "An update on the prevalence of cerebral palsy: A systematic review and meta-analysis," *Develop. Med. Child Neurol.*, vol. 55, no. 6, pp. 509–519, Jun. 2013.

[42] C. Peyton and C. Einspieler, "General movements: A behavioral biomarker of later motor and cognitive dysfunction in NICU graduates," *Pediatric Ann.*, vol. 47, no. 4, pp. e159–e164, Apr. 2018.

[43] K. Raghuram *et al.*, "Automated movement recognition to predict motor impairment in high-risk infants: A systematic review of diagnostic test accuracy and meta-analysis," *Develop. Med. Child Neurol.*, vol. 63, no. 6, pp. 637–648, Jun. 2021.

[44] K. Raghuram *et al.*, "Automated movement analysis to predict motor impairment in preterm infants: A retrospective study," *J. Perinatol.*, vol. 39, no. 10, pp. 1362–1369, Oct. 2019.

[45] H. Rahmati, H. Martens, O. M. Aamo, O. Stavdahl, R. Stoen, and L. Adde, "Frequency analysis and feature reduction method for prediction of cerebral palsy in young infants," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 11, pp. 1225–1234, Nov. 2016.

[46] S. Reich *et al.*, "Novel AI driven approach to classify infant motor functions," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Dec. 2021.

[47] S. M. Reid *et al.*, "Temporal trends in cerebral palsy by impairment severity and birth gestation," *Develop. Med. Child Neurol.*, vol. 58, pp. 25–35, Feb. 2016.

[48] E. Ricci, C. Einspieler, and A. K. Craig, "Feasibility of using the general movements assessment of infants in the united states," *Phys. Occupational Therapy Pediatrics*, vol. 38, no. 3, pp. 269–279, May 2018.

[49] P. Rosenbaum *et al.*, "A report: The definition and classification of cerebral palsy April 2006," *Develop. Med. Child Neurol. Suppl.*, vol. 109, pp. 8–14, Mar. 2007.

[50] W. Rueangsirarak, J. Zhang, N. Aslam, E. S. L. Ho, and H. P. H. Shum, "Automatic musculoskeletal and neurological disorder diagnosis with relative joint displacement from human gait," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 12, pp. 2387–2396, Dec. 2018.

[51] D. Sakkos, K. D. Mccay, C. Marcroft, N. D. Embleton, S. Chattopadhyay, and E. S. L. Ho, "Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy," *IEEE Access*, vol. 9, pp. 94281–94292, 2021.

[52] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf

[53] A. Stahl, C. Schellewald, Ø. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerød, "An optical flow-based method to predict infantile cerebral palsy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 20, no. 4, pp. 605–614, Jul. 2012.

[54] P. Turaga, R. Chellappa, and A. Veeraraghavan, "Advances in video-based human activity analysis: Challenges and approaches," in *Advances in Computers*, vol. 80, M. V. Zelkowitz, Ed. Amsterdam, The Netherlands: Elsevier, 2010, pp. 237–290.

[55] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.

[56] Q. Wu, G. Xu, F. Wei, L. Chen, and S. Zhang, "RGB-D videos-based early prediction of infant cerebral palsy via general movements complexity," *IEEE Access*, vol. 9, pp. 42314–42324, 2021.

[57] M. Zhu, Q. Men, E. S. L. Ho, H. Leung, and H. P. H. Shum, "Interpreting deep learning based cerebral palsy prediction with channel attention," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.

[58] D. Zlatanovic *et al.*, "The importance of the Prechtl method for ultra-early prediction of neurological abnormalities in newborns and infants," *Appl. Mech. Mater.*, vol. 58, no. 3, pp. 111–115, Sep. 2019.