Received May 31, 2021, accepted June 24, 2021, date of publication June 29, 2021, date of current version July 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093469

Identification of Abnormal Movements in Infants: A Deep Neural Network for Body Part-Based Prediction of Cerebral Palsy

DIMITRIOS SAKKOS[1], KEVIN D. MCCAY[1], CLAIRE MARCROFT[2], NICHOLAS D. EMBLETON[2], SAMIRAN CHATTOPADHYAY[3], (Senior Member, IEEE), AND EDMOND S. L. HO[1]

[1]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K.
[2]Newcastle Neonatal Services, Newcastle upon Tyne Hospitals NHS Foundation Trust, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, U.K.
[3]Department of Information Technology, Jadavpur University, Kolkata 700054, India

Corresponding author: Edmond S. L. Ho (e.ho@northumbria.ac.uk)

This work was supported in part by the Royal Society under Grant IES \R1\191147.

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

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 these manual methods can be laborious. The prospect of automating these processes is seen as key in advancing this field of study. In our previous works, we examined 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 new deep learning framework for this classification task. We also propose a visualization framework which identifies body-parts with the greatest contribution towards a classification decision. The inclusion of a visualization framework is an important step towards automation as it helps make the decisions made by the machine learning framework interpretable. We directly compare the proposed framework's classification with several other methods from the literature using two independent datasets. Our experimental results show that the proposed method performs more consistently and more robustly than our previous pose-based techniques as well as other features from related works in this setting. We also find that our visualization framework helps provide greater interpretability, enhancing the likelihood of the adoption of these technologies within the medical domain.

INDEX TERMS Cerebral palsy, deep learning, early diagnosis, explainable AI, general movements assessment, interpretable AI, machine learning, medical visualization, motion analysis, skeletal pose.

I. INTRODUCTION

The process of automating the recognition, analysis and reconstruction of complicated motion, such as human activity, has been an area of interest for researchers in many varied fields, due to its inherent ability to streamline intensive manual processes. [5]. It has seen widespread use in surveillance, virtual reality, intelligent monitoring and content based video indexing [48]. In our previous works [32], [33], we proposed different methods by which we could examine the feasibility of applying these technologies to the healthcare domain, specifically to aid with the early prediction of cerebral palsy (CP) in infants.

CP is an umbrella term used to describe a group of lifelong neurological conditions which cause movement difficulties. These movement difficulties typically affect mobility, posture and coordination, but can also cause problems with speech articulation, swallowing, vision, and can contribute towards a reduced ability to learn new skills. The severity of these symptoms can vary quite significantly, with some individuals presenting very minor symptoms whilst others may be severely disabled [20].

The associate editor coordinating the review of this manuscript and approving it for publication was Imran Sarwar Bajwa.

CP is attributed to non-progressive damage to the brain in early infancy [9], [42] and is one of the most common physical disabilities in childhood, with an overall prevalence of 2.11 per 1000 live births [37]. Studies have also found that there is a significant link between CP and infants born prematurely, with the prevalence of CP in infants born very preterm (28-32 weeks gestation) being 32.4 per 1000 surviving neonatal births, and for infants born extremely preterm (less than 28 weeks gestation) 70.6 per 1000 births [41]. Studies also suggest that whilst improvements to neonatal care have led to a decline in infant mortality rates, they have also contributed to an increase in the frequency and severity of CP [38].

CP is a lifelong condition where earlier diagnosis improves outcomes by facilitating faster access to physiotherapy support and enhancing parental understanding. This results in better physical outcomes for the child over the life-course as well as providing better support for families. In order to provide opportunities for the best possible outcome for an infant's development, early diagnosis of CP is considered essential. The early diagnosis of CP can however be difficult, with confirmed diagnosis often not being made until 18 months of age, or potentially later for individuals with mild symptoms [31]. Currently, tests which identify the emerging signs of CP typically evaluate the quality, complexity and spontaneity of an infant's movements at a specific window in their development, one such test is the General Movements Assessment (GMA) [15].

The GMA evaluates infant movement by observing the presence of ''Fidgety Movements'' (FMs), which are observable from 3 to 5 months post term [39]. In a typically developing infant FMs have a similar prevalence and appearance, subsequently allowing for abnormal FM patterns to be identified and classified [14]. The GMA has produced some excellent results in detecting CP [35], indeed in a recent review study, [10] reports that the GMA produces more reliable results than other methods such as cranial ultrasound and neurological examination.

In practice, the challenges of applying these assessments relates to the availability of appropriately trained and skilled clinicians. These clinicians require considerable additional training and as such, tests are currently only carried out on infants at high risk of developing CP [8]. These tests are based around the gestalt visual perception of movement [16], and are therefore highly subjective, lacking discernible quantitative diagnostic features. These clinical tests are also heavily reliant upon the infant being in a suitable behavioural state [19], making them potentially very time-consuming, which can subsequently lead to observer fatigue.

In order to address the issues found in manual clinical assessment, several works [2], [32], [33], [36], [43] have attempted to automate the process of GMA for the prediction of CP using machine-learning. Not only would an automated system have the potential to reduce the time and cost associated with current manual clinical assessments, but it could also assist clinicians in making an earlier and more confident diagnosis by providing interpretable, quantitative information about why a prediction of CP has been made.

One of the main issues with using machine-learning approaches in the medical domain is the problem of interpretable AI. Models are often seen as 'black boxes' in which the underlying structures can be difficult to understand. There is an increasing requirement for the mechanisms behind why systems are making decisions to be transparent, understandable and explainable [24]. As such, we propose a new motion classification framework, which takes an RGB video as the input and analyzes the movement of individual body parts to determine if FMs are present (FM+) or absent (FM−), subsequently identifying normal or abnormal general movements from segments of the sequence. To make our proposed framework fully interpretable, an important aspect is the automatically generated visualization capable of relaying pertinent information to the assessor. This visualization highlights the segmented body parts that are showing movement abnormalities, and are subsequently providing the most significant contribution towards the final classification result.

Another issue found in the existing methods [2], [32], [33], [36], [43] is the lack of comparisons of performance between different approaches in the literature. Moreover, due to the sensitive nature of the video data recorded, the datasets used in the previous works are not available to the public. As a result, it is difficult for researchers to have a fair comparison and evaluation on the performance of different methods due to the unavailability of benchmark datasets. In this work, we compare the classification performance with several state-of-the-art CP prediction frameworks on 2 datasets, details of which are discussed in Section III-A. The details of our proposed classification and visualisation frameworks are discussed in Sections III and IV, and our evaluation is discussed in Section V. Our hope is that this contribution will aid in the adoption of such technologies in this domain, through accurate, quantifiable and explainable results.

The contributions of this work can be summarized as:

- A new deep neural network based classification framework is proposed for the automated prediction of Cerebral Palsy based upon body-part movements extracted from RGB videos.
- A visualization feature is proposed to highlight the contribution of each body segment towards the Cerebral Palsy prediction in the video to improve the model interpretability.
- A challenging new dataset was constructed using real patient data gathered as part of routine clinical care. This dataset reflects the intra-class variance and associated complexity found in carrying out manual assessments in a real-world clinical environment.

II. RELATED WORKS

In this section, we discuss several works related to the automation of the GMA, we consider the techniques used and how our proposed system might take advantage of and build upon these existing methods. In a recent study [31] discussed

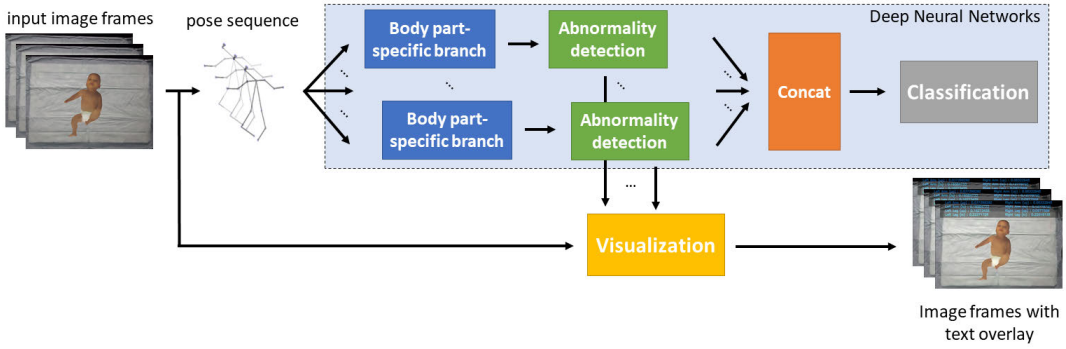


FIGURE 1. The overview of the proposed framework.

the use of several technologies which analyse recorded movement data for this purpose. They suggest that the adoption of these technologies could potentially help with the prediction of motor impairment in infants. The study also highlighted the need for the early identification of those most at risk, and proposed that the application of these technologies within paediatric practice could also contribute towards a better understanding infant neurological development.

Several studies have been carried out which attempt to assess the viability of automating assessments to predict motor impairment based upon observed motion quality, such as the GMA, using computer vision-based approaches. The first examples of this can be found in the works by Adde *et al.* [1], [2], [4], who developed a per frame background subtraction method for analysis. A simple representation is created which calculates the difference between two frames in a video sequence. A point value per pixel of 0 or 1 is then assigned which represents the presence of movement. Relevant features were extracted from this data and used for classification. Whilst reasonable results were obtained with this method, greater robustness and understanding of the motion associated with FMs was desired, as such researchers started implementing more advanced motion assessment techniques, such as optical flow.

In [43], an optical flow-based method was produced which used statistical analysis and pattern recognition to assess the infant's spontaneous movements. Wavelet frequency analysis was then used to evaluate the time-dependant trajectory signals found in the optical flow data. Similarly, Orlandi *et al.* [36] implemented large displacement optical flow (LDOF) to track infant body movements and acquire velocities. The displacement of each pixel was calculated prior to extracting features for a binary classification, to determine normal or abnormal general movements (GMs). Ihlen *et al.* [25] also made use of an LDOF model to track infant movements through a pixelwise representation. The displacement was manually annotated to determine the likelihood of cerebral palsy based upon the proportion of CP risk-related movements. This was then classified using statistical analysis of the data. Rahmati *et al.* [40] also used optical flow, proposing the use of features derived from frequency analysis to classify infant motion into one of two groups. Using video sequences as input, a motion segmentation algorithm was used to extract motion data from each limb. Their proposed feature selection method, which determines features with significant predictive ability, is then used prior to classification.

Each of the methods discussed make use of traditional machine learning classification algorithms, however, in recent years researchers have started to develop new methods of classification for motion analysis. These methods are typically more robust, accurate and reliable than previous methods, incorporating deep learning frameworks for this classification task.

Some leading examples of how deep learning frameworks have been implemented for action recognition, human motion analysis and classification can be found in works exploring pose estimation [12], [18], [21], [29], [44] and part-based segmentation [18], [45], [47]. In these works researchers have proposed alternate methods to model human shape and motion as a means of overcoming the limitations found in other methods, such as optical flow. However, these deep learning based approaches are largely yet to be implemented in the field of automated CP prediction in infants due to the difficulty in model interpretability and the availability of suitable data. Our proposed system takes advantage of these improvements in motion analysis performance by incorporating state-of-the-art methods in this classification task, details of which are discussed in Section III. However, since we also incorporate a visualization segment, our system also benefits from greater contextual interpretability than typical deep learning frameworks. This allows our system to retain the explainability found in earlier related works, whilst simultaneously capitalising on the performance improvements offered by employing deep learning architectures.

III. METHODOLOGY-OVERVIEW AND DATA PRE-PROCESSING

In this section we present a general overview of the proposed framework, illustrated in Figure 1, and the data pre-processing tasks. Given an image sequence, extracted from top down video recordings of infant general movements,

the 2D skeletal pose is detected on a per frame basis using OpenPose [11]. The input of the proposed framework is the extracted 2D skeletal pose sequence. For the skeletal motion data, each pose is divided into different body-parts and each body-part sequence is processed by a specific branch (Section IV-A) to learn a part-specific spatio-temporal representation. To evaluate the importance of each body-part in contributing to the final classification result, deep supervision is added to each part-specific stream. Finally, the outputs from all the individual body-part streams are concatenated and fed to the classifier (Section IV-C) to predict the final label of the image sequence. In addition, the attention information (Section IV-B) which indicates the presence of abnormalities in the movement of each body part is combined with the video as a visualization (Section IV-D) of the automated prediction process.

In the rest of this section, we introduce the datasets used in this study in Section III-A, and we explain the 2D skeletal pose extraction and associated pre-processing required in Section III-B.

A. DATASETS

We make use of two different GMA datasets, a synthetically generated dataset (MINI-RGBD) and a real-world dataset (RVI-25). Here, we discuss the details of these datasets and their implementation in our proposed fidgety movement identification and visualization pipeline.

1) MINI-RGBD

Due to the sensitive nature of the video recordings required for the GMA, the Moving INfants In RGB-D (MINI-RGBD) [23] synthetic dataset was produced and made publicly available. The MINI-RGBD dataset consists of 12 synthetically generated videos (640×480 resolution @ 25 FPS), each 1000 frames long, showing the movements of infants lying in a supine position. These videos were created by recording infants in a clinical setting and mapping the corresponding real-world movements to synthetically generated 3D models of infants. This approach maintains the anonymity of the infants whilst fully and accurately representing their motion characteristics. We make use of this synthetic dataset and label each of the video sequences as 'normal' (FM+) or 'abnormal' (FM−) based upon the GMA. The data labelling was carried out by assessors highly experienced in the clinical application of the GMA.

2) RVI-25

The second dataset (RVI-25), consists of 25 videos of 25 different infants, recorded as part of routine clinical care at the NHS Royal Victoria Infirmary, Newcastle upon Tyne, UK. All necessary ethical approvals were provided by the Research Ethics Committee (REC), HRA and HCRW. Written informed consent was also obtained from the parents/legal guardians of all participating infants. The video recordings were carried out using a handheld 2D RGB video camera (Sony DSC-RX100 with 1.0-Type Sensor, 28-100 mm F1.8-4.9 Zeiss Lens, recording in 1920×1080 resolution @ 25 FPS) during active wakefulness, filmed from above with the infant lying in a supine position. The video was filmed in this way so as to fully simulate a real-world recording environment, to assess the efficacy of the extracted features, and enable future analysis on patient footage within both the clinical setting and remotely. The length of each video varies between 1 and 5 minutes, for our implementation we extracted individual sequences 1000 frames in length, our ultimate aim being to establish if FMs can be detected in short sequences to enable the spatio-temporal identification of risk related movements. Each of these sequences was screened and assessed for the presence of FMs (FM+ or FM−), according to the GMA, by an experienced paediatric physiotherapy team.

B. DATA PRE-PROCESSING

Whilst the infant movement data is captured as RGB video, directly analyzing videos is a challenging task since a wide range of factors contributes towards the intra-class variations, including illumination, the background of the video, the appearance of the infant (body shape, skin color, with or without clothing), etc. In contrast, encouraging results have been reported [32], [33] in infant movement analysis based on skeletal pose features extracted from videos. In the proposed framework, 2D skeletal pose sequences are used as the input for the analysis of infant general movements.

1) POSE ESTIMATION FROM VIDEO

For pose estimation, the locations of body parts (e.g. joints) can be detected from an image. In particular, OpenPose [11] is one of the top-performing approaches proposed in recent years. OpenPose is based on Part Affinity Fields (PAFs), which learn the association between body parts and their appearance in the image. Such an approach is also referred to as a 'bottom-up' approach that recognizes lower level features (e.g. body parts) first, in order to reconstruct the higher level skeletal posture. An example of the skeletal pose extracted using OpenPose is shown in Figure 2.

In this study, the official OpenPose implementation (https://github.com/CMU-Perceptual-Computing-Lab/openpose) is used for extracting the 2D locations of the joints from the video. Specifically, each video is converted into a sequence of images and a skeletal pose is extracted from each image. For each posture, 18 keypoints including body joint locations and facial landmarks are detected. An example is shown in Figure 2. Each keypoint contains the x and y coordinates of the joint location within the image. In this work, 14 joints, including *head, neck, left and right shoulders, left and right elbows, left and right wrists, left and right hips, left and right knees, and left and right ankles* were used.

2) AUTOMATIC DATA CORRECTION

The accuracy of joint location prediction can, however, be affected by factors such as self-occlusion of body parts. To alleviate this problem, an automatic data correction

FIGURE 2. An example of the extracted pose estimation using OpenPose [11] on the MINI-RGBD [23] dataset.

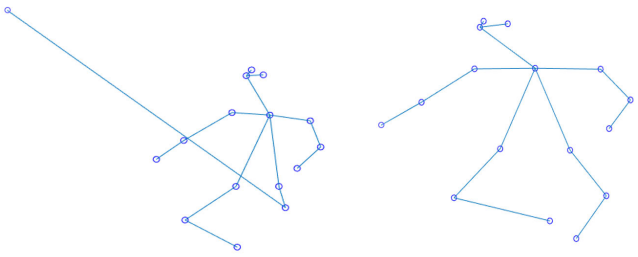


FIGURE 3. An example of the automated pose data correction approach presented in [33]. Note that for this particular frame one of the joints in the original pose estimation has defaulted to position 0,0 due to occlusion, this joint has subsequently been adjusted in the corrected version based upon the calculated confidence score. The corrected pose also shows the result of the data normalization process.

approach, presented in [33], is used. An example of this is shown in Figure 3. Specifically, OpenPose returns a confidence score associated with each predicted joint location. Since different confidence score distributions are obtained from different videos, due to the different movements as well as environmental conditions (such as lighting, video quality, etc), we adaptively adjust the threshold of the confidence score to decide whether the predicted joint location is 'usable' or correction is required. In particular, we compute the threshold value t_i for joint i by

$$t_i = (\frac{1}{n}\sum_{j=1}^{n} c_{i,j}) \times 90\% \quad (1)$$

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. We follow McCay *et al.* [33] on multiplying the averaged confidence value by 10% as the threshold.

Next, the trajectory of each joint is computed separately by curve fitting based on the joints with confidence scores which are above the threshold. In doing so, the location of the joint with a confidence score below the threshold at a frame will be estimated by the locations in the neighbouring frames with a higher confidence score. This aligns with the observation that human motion is continuous and the videos are captured at a high frame rate (25 FPS or above). As a result, the changes of the joint locations over time should be small and a curve function can approximate the joint trajectory over time. Among a wide range of curve fitting functions, the modified Akima interpolation [7] is selected in our work, since using this spline function can effectively avoid the overshooting issues found in other spline functions. This results in a more natural interpolated trajectory, closer to the original signal. The joint locations with a confidence score above the threshold will then be used as the control point X_j as the input of the modified Akima interpolation.

3) DATA NORMALIZATION

The joint locations returned from OpenPose are presented as the x and y coordinates on the image. Since differences in body size may impact upon the magnitude of body movement, subsequently affecting classification accuracy, all of the extracted infant pose data is scaled to a standardized height. We do this by computing the height of the infant from the skeletal pose. The height can be estimated by the sum of the lengths of the following body segments: *lower leg, upper leg, hip-to-neck and head* (refer to Figure 2). In addition to the scaling factor, the orientation of the infant also affects the performance of the machine learning process. As such, we transform the extracted pose sequence so that the infant pose data is vertically aligned and centred in each frame, as illustrated in 3. We do this by computing the acute angle between the medial axis of the torso, which can be represented by a straight line between the middle of the hip joints (i.e. left hip and right hip) and the neck joint, and a vertical line in the coordinates system. Each posture in the motion sequence is then rotated and centred according to the computed acute angle. We apply this to all of the extracted posture sequences, meaning they are normalized and ready for analysis. Finally, before we feed the data to the model, we subtract the coordinates of each frame to these of the first frame.

IV. METHODOLOGY-THE PROPOSED FRAMEWORK

In this section, the details of the our proposed framework will be given in Section IV-A to IV-C, which include the body part specific branch (Section IV-A) with abnormality detection (Section IV-B) and the classification (Section IV-C). The proposed visualization module will be presented in Section IV-D.

A. PART-BASED MOVEMENT MODELLING USING CNN AND LSTM

The GMA assesses the overall quality of infant movement at a specific window in their development. One of the most important criteria assessed in this process is whether normal FMs are present or absent [3]. These FMs are defined by Prechtl as a continuous stream of movements in all directions by multiple body parts, with moderate speed and variable acceleration [39]. Inspired by this, we propose a system which models the movement of each body part

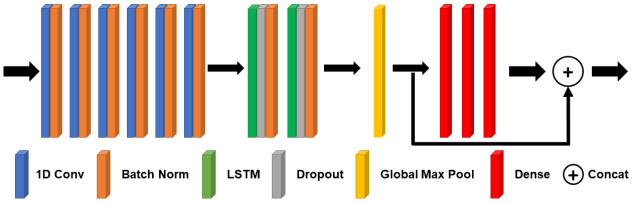


FIGURE 4. The network architecture of the body part-specific stream.

separately, in order to analyze how the movement from each part contributes towards the final classification decision. Ultimately, our framework determines whether FMs are absent (FM−) or present (FM+), and consequently classifies whether the movement of the infant can be considered normal or abnormal.

Specifically, we model the trajectories of 8 selected body joints individually, including right elbow, right hand, left elbow, left hand, right knee, right ankle, left knee and left ankle. The movement trajectory of each selected body joint is essentially a spatio-temporal motion representation, as it contains the joint location information over time. The trajectory of a joint j extracted from a video with n frames is represented by:

$$P_j = \begin{bmatrix} p_{j,x,1} & \cdots & p_{j,x,n} \\ p_{j,y,1} & \cdots & p_{j,y,n} \end{bmatrix} \tag{2}$$

where $p_{j,x,i}$ and $p_{j,y,i}$ are the $x-$ and $y-$ coordinates of joint j at frame i.

1) SPATIAL MODELLING USING 1D CNNs

Next, the joint trajectory is encoded spatially. When handling signals with relatively low dimensionality, 1D convolutional neural networks (CNNs) is preferred over 2D CNNs since 1D CNNs are more efficient with a shallower architecture for learning challenging 1D signals [26]. This property is crucial for analyzing infant movements since only limited data sample are available and it will be difficult to train a deep neural network which usually requires a huge amount of data. As a result, we propose using 1D CNN to model each body part spatially and the network architecture of each body part-specific stream is illustrated in Figure 4. Specifically, 5 mini blocks are used for spatial modelling. In each mini block, the input is passed to a 1D convolution layer followed by a batch normalization layer. The purpose of passing the output of the 1D convolution layer into the batch normalization layer is to reduce the covariance shift, and results in a more stable and efficient training process. By stacking the mini blocks 5 times we can extract semantically rich features of lower spatial dimensionality, as it is gradually decreased through strided convolutions.

2) TEMPORAL MODELLING USING LSTM

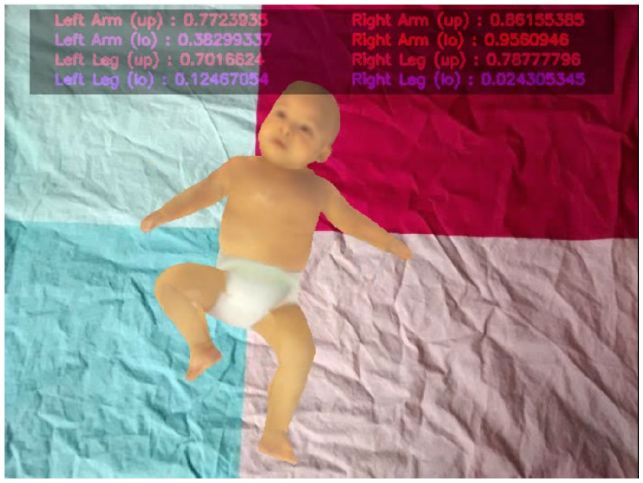
To model the movement of each body part temporally, Long Short Term Memory (LSTM) network architecture is used as shown in Figure 4. Here, we use 2 mini blocks for this purpose, and each mini block includes a LSTM layer followed by a dropout layer and finally a batch normalization layer. The effectiveness of using LSTM for modelling temporal sequence has been demonstrated in a wide range of applications [22]. Furthermore, LSTM is preferred over a temporally-extended CNN for two reasons. Firstly, the LSTM network is inherently superior to handling temporal information, as it is equipped with special modules such as the forget gate, whereas CNNs treat the temporal dimension as a third spatial dimension [30]. Secondly, the LSTM is capable of modelling the entire sequence globally and can focus on multiple different segments. On the other hand, CNNs extract local features and require a large receptive field, usually focusing on fewer segments [30]. While LSTMs are usually more computationally expensive, this is less relevant in this work as our model is quite small both in depth and in width. To prevent overfitting, the output of the LSTM layer is fed to a dropout layer to improve the generalisation of the trained model. Finally, a batch normalization layer is added to improve the stability and efficiency of the network. We repeat this mini block architecture twice as a deeper network improves the learning capability.

B. BODY-PART ABNORMALITY DETECTION

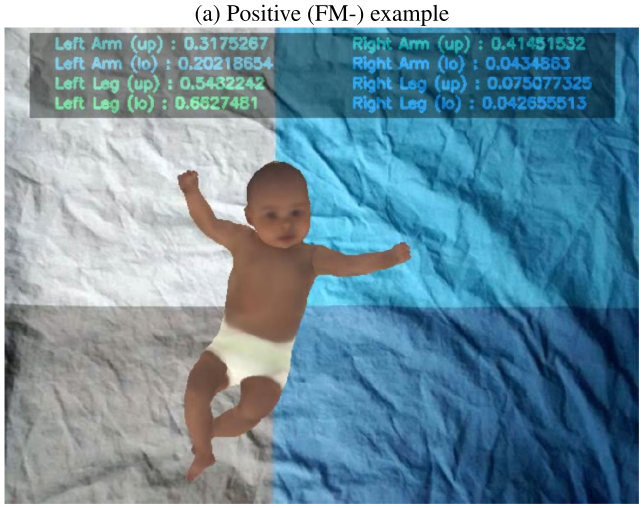
To make the automated FM identification system more interpretable, we feed the output of each body part-specific branch to a classifier which is trained using the annotation (i.e. FM+ or FM−) of the video. This design allows the system to provide additional information on the importance of each stream (i.e. body part) in determining the abnormality of the body movement of infants. The importance of these individual body part is used to highlight the body parts with the highest classification contribution to inform clinicians where they should pay attention to for further analysis.

On that end, after the spatio-temporal modelling of movement of the body part, the CNN-based architecture produces a high dimensional output. To facilitate the classification learning process, a down-scaling process is usually included. In particular, global max pooling is used in our framework, to highlight the abnormalities encoded in the embedding vector. Two fully connected layers are appended to encode abnormality information, followed by the classification layer. Before feeding the output to the classifier, we concatenate the predicted class to the deep embedding, so the classifier has access to the stream's prediction.

The output of each stream will be a contribution score (i.e. a scalar value) which indicates how much the body-part contributes to the classification decision. In addition to the overall classification results (i.e. FM+ or FM−) (Section IV-C, the contribution score will also be provided as additional feedback to use to clinicians to understand why the decision was made. An example is illustrated in Figure 5. Based on the contribution score, we use purple (0) to red (1) as the color range for positive class, and the color range from blue (0) to green (1) for negative class.



(a) Positive (FM-) example



(b) Negative (FM+) example

FIGURE 5. Examples of the video generated by our visualization module.

C. CLASSIFICATION

Finally, the outputs from all stream will be merged and fed to a classification network illustrated in Figure 6 to obtain the overall result. Specifically, after concatenating the stream outputs, an instance normalisation is performed to eliminate the scale differences between streams. Then, a 1D convolution with 100 filters is applied to learn the best fusion weights. A dropout layer randomly drops 50% of its neurons, to avoid overfitting the training data. Finally, the output is fed to a dense layer which is followed by a sigmoid layer for classification. For the RVI-25 dataset, which is larger and more complex, we add two more fully connected layers of 50 and 150 neurons each.

D. VISUALIZATION AS AN INTERPRETABLE AUTOMATED ASSESSMENT

In order to make our proposed framework more interpretable, we include a visualization module which highlights the body parts that are contributing to the classification decision on the image frames. While deep learning-based frameworks

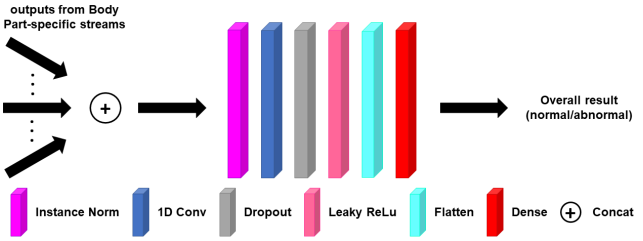


FIGURE 6. The network architecture of the classification.

obtained excellent performance in a wide range of visual understanding tasks, most of the existing architectures can be considered as black-box approaches. For example, most of the classification frameworks only output the predicted label without specifying how the conclusion is being drawn. Whilst this is acceptable in typical computer vision tasks, it is less preferable in healthcare and medical applications, since it is essential for the clinicians to verify the prediction as well.

To make the proposed framework more interpretable, the body part abnormality detection explained in Section IV-B is used to evaluate the contribution from each stream towards the overall classification. Specifically, each body part stream will be associated with a contribution score which is a normalized value between 0 and 1. Note that the score is computed from the spatio-temporal body part representation, and this information is displayed as video overlay. The contribution score will then be used to determine the color intensity of the fonts. For positive class, we use purple (0) to red (1) as the color range. For negative class, the color range from blue (0) to green (1) is being used. An example of the visualization result is illustrated in Figure 5.

V. EVALUATION

In this section, we evaluate the performance of our proposed framework quantitatively and qualitatively. We compare the classification performance of our method with several other baseline methods proposed in the literature. The baselines used here represent a selection of state-of-the-art methods in the task of identifying fidgety movements for CP diagnosis. They provide robust performance backed up by several publications and associated implementations, as well as source code availability, making them particularly useful for our evaluation.

We evaluated each of the feature extraction and motion classification methods using two separate datasets, each consisting of videos showing infant movement characteristics associated with the GMA (details of each dataset are discussed in Section III-A). For fair comparison, all of the features extracted using the baseline methods were classified using the Support Vector Machine (SVM), Decision Tree (Tree), k-NN (k = 1 and k = 3), Linear Discriminant Analysis (LDA), Ensemble, and Logistic Regression classification algorithms, with the best results for each feature and dataset reported in Table 1 and Table 2.

To our knowledge, a comparison of the different proposed methods has not been carried out to quantitatively evaluate the effectiveness of each extracted feature on shared datasets. We also examine the effectiveness of the visualization portion of our framework through qualitative assessment. Our results and observations of the classification performance comparison are discussed in Section V-C, and our evaluation of the visualization framework is discussed in Section V-D.

A. EXPERIMENTAL SETTINGS

Both datasets, but the synthetic in particular, are very limited in size. Therefore, overfitting is a major problem which had to be addressed. The following measures have been put in place:

- Dropout layers: As discussed in Sections IV-A2, IV-C, dropout layers are used to improve generalisation.
- Data augmentation: Each training sample is augmented with 80% probability with the following methods:
  - -- Random scaling: A random value is selected from the interval [0.35, 1.65] to downscale or upscale the training sample.
  - -- Noise addition: A small value, up to a third standard deviation is added to each coordinate of the training data.
  - -- Sign inversion: The sign of the input values is changed by 50% probability.
  - -- Motion reverse: The motion is reversed by 50% probability.

In addition, another challenging problem that is inherent to the datasets is class imbalance. The synthetic dataset is biased to the negative class, as it has twice as many videos of that class. The imbalance is even greater for the real dataset, with negative videos being more than three times as many. Therefore, during training, we sample from the positive or negative class with equal probability 50%, rather than selecting a train video uniformly. As a result, each batch has approximately the same number of positive and negative classes. This ensures the model will be unbiased towards the two classes.

For the leave-one-out cross validation in this work, every sample in the dataset will be selected as the testing sample while the rest are used as the training and validation sets. The results reported in Tables 1 and 2 are the averaged accuracy, specificity and sensitivity obtained from all samples in each dataset. Due to the limited number of samples in the datasets, only two samples (1 positive and 1 negative) are selected as the validation set and the remaining samples (excluding the testing sample) are used as the training set. The selection of validation set samples follows the procedures below:

1) randomly select a sample from each of the categories (i.e. positive and negative)
2) start training and predict the class labels on the validation set
3) go back to step 1 again if the validation accuracy is not 100%

This validation set selection approach keeps improving the validation accuracy while maintaining the heuristic nature of sample selection. We use 1000 frames in all training and validation samples.

Regarding the training parameters, a batch size of 20 was used. The Adam optimizer was selected with a starting learning rate of 0.001. We train for 20 epochs of 30 iterations and use a cosine annealing learning rate decay with warm restarts [27]. The cycle length is 6 epochs and the minimum learning rate is 5e-6. Our proposed framework is implemented on Keras with Tensorflow. All the experiments are executed on a computer with an NVidia Titan Xp GPU.

TABLE 1. Classification accuracy comparison between our proposed framework and baseline methods on the MINI-RGBD [23] dataset with a Leave-One-Out data split. For clarity, only the best results from the classification algorithms discussed in Section V are reported.

Feature	Bins	Classifier	AC	SE	SP	PR	F1
CX_{mean} [4]		Log. Reg.	0.667	0.500	0.750	0.500	0.500
CX_{SD} [4]		3-NN	0.667	0.750	0.625	0.500	0.600
CY_{mean} [4]		SVM	0.667	0.000	**1.000**	-	-
CY_{SD} [4]		3-NN	0.750	0.750	0.750	0.600	0.667
C_{SD} [1]		Ensemble	0.750	0.500	0.875	0.667	0.571
Q_{mean} [1]		1-NN	0.500	0.250	0.625	0.250	0.250
Q_{SD} [1]		3-NN	0.750	0.750	0.750	0.600	0.667
CPP [1]		Ensemble	0.667	0.500	0.750	0.500	0.500
AMD [43]	32	Ensemble	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Freq. [43]	8	LDA	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Wavelet [43]	128	LDA	0.833	0.750	0.875	0.750	0.750
Wu et al. [46]			0.917	1.000	0.875	0.800	0.889
HOJO2D [32]	8	LDA	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HOJO2D [32]	16	LDA	0.750	0.750	0.750	0.600	0.667
HOJD2D [32]	8	Ensemble	0.833	0.750	0.875	0.750	0.750
HOJD2D [32]	16	LDA	0.750	0.500	0.875	0.750	0.750
HOJO+JD [32]	8	LDA	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HOJO+JD [32]	16	LDA	0.833	0.750	0.875	0.750	0.750
Our Proposed Method			**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

TABLE 2. Classification accuracy comparison between our proposed framework and baseline methods on the RVI-25 dataset with a Leave-One-Out data split. For clarity, only the best results from the classification algorithms discussed in Section V are reported.

Feature	Bins	Classifier	AC	SE	SP	PR	F1
CX_{mean} [4]		DT	0.680	0.333	0.790	0.333	0.333
CX_{SD} [4]		DT	0.760	0.333	0.895	0.500	0.400
CY_{mean} [4]		Ensemble	0.640	0.167	0.790	0.200	0.182
CY_{SD} [4]		Ensemble	0.760	0.333	0.895	0.500	0.400
C_{SD} [1]		SVM	0.760	0.000	**1.000**	-	-
Q_{mean} [1]		Ensemble	0.720	0.333	0.842	0.400	0.364
Q_{SD} [1]		Ensemble	0.720	0.333	0.842	0.400	0.364
CPP [1]		1-NN	0.720	0.167	0.895	0.333	0.222
AMD [43]	32	LDA	0.840	**0.833**	0.843	0.625	0.714
Freq. [43]	64	LDA	0.720	**0.833**	0.684	0.455	0.588
Wavelet [43]	128	LDA	0.760	0.333	0.895	0.500	0.400
HOJO2D [32]	8	SVM	0.880	0.500	**1.000**	**1.000**	0.667
HOJO2D [32]	16	SVM	0.800	0.500	0.895	0.600	0.545
HOJD2D [32]	8	LDA	0.880	0.667	0.947	0.800	0.727
HOJD2D [32]	16	Log. Reg.	0.680	0.500	0.737	0.375	0.429
HOJO+JD [32]	8	SVM	0.800	0.500	0.895	0.600	0.545
HOJO+JD [32]	16	SVM	0.880	0.500	**1.000**	**1.000**	0.667
Our Proposed Method			**0.920**	**0.833**	0.947	0.833	**0.833**

B. ABLATION STUDY

To justify the design of the proposed framework, an ablation test was conducted to evaluate how different settings

TABLE 3. An ablation study on using different number of joints in the proposed framework for CP prediction on the MINI-RGBD [23] and RVI-25 datasets.

Dataset	# of joints	AC	SE	SP	PR	F1
MINI-RGBD [23]	8	1.000	1.000	1.000	1.000	1.000
	14	0.917	0.750	1.000	1.000	0.857
RVI-25	8	0.920	0.833	0.947	0.833	0.833
	14	0.880	0.500	1.000	1.000	0.667

affect the system performance in CP prediction. In particular, the selection of body parts to be analyzed is evaluated. In our pilot study [32], [33], using the motion data extracted from the limbs resulted in a better classification performance. We compared the performance of the proposed framework using all joints (14 joints) and only the 8 joints extracted from the limbs. The results are presented in Table 3. The results confirm that using limb-based 8 joint variant led to the best performance on both datasets. One possible reason for the sub-optimal performance obtained using all joints could be related to the complexity of the network architecture. In the future, we will explore the feasibility of increasing the number of layers of the classification sub-network (Figure 6). However, we expect that more training data will be required as the network architecture becomes more complex.

C. CLASSIFICATION RESULTS AND DISCUSSION

We compare our method with several other previously proposed methods, which form the baselines for our comparison. These baselines ([1], [4], [32], and [43]) generally represent the current state-of-the-art approaches currently in use for this classification task. In line with several related works from the literature [6], [14], [17], [28], we report the best average classification accuracy, sensitivity, specificity, precision and f1-scores for each method, using a leave-one-out data split.

$$SE = \frac{TP}{TP + FN} \tag{3}$$

$$SP = \frac{TN}{TN + FP} \tag{4}$$

$$AC = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

$$PR = \frac{TP}{TP + FP} \tag{6}$$

$$F1 = \frac{2 * TP}{2 * TP + FN + FP} \tag{7}$$

In our evaluation, true positive (TP) represents cases in which impaired infants are correctly diagnosed as impaired, true negative (TN) represents unimpaired infants correctly identified as unimpaired, false positive (FP) represents unimpaired infants incorrectly identified as impaired, and false negative (FN) represents impaired infants incorrectly identified as unimpaired. As such, the sensitivity (SE) can be defined as the percentage of positive classifications amongst the positive population of the dataset, the specificity (SP) can be defined as the percentage of negative classifications amongst the negative population of the dataset and the accuracy (AC) can be defined as the percentage of correctly classified instances. Table 1 shows the results using the MINI-RGBD dataset, and Table 2 shows the results using the RVI-25 dataset.

When using the MINI-RGBD dataset, we observe that the results based upon the Centroid of Motion features (CX_{mean}, CX_{SD}, CY_{mean}, CY_{SD}, and C_{SD}) and the Quantity of Motion features (Q_{mean} and Q_{SD}) generally do not perform as well in each of the evaluated metrics. We see a relatively low accuracy (0.5000 to 0.7500), and sensitivity (0.5000 to 0.7500), and a highly varied specificity (0.6250 to 1.000). This is also true of the combined features used in the CPP feature set, with similar patterns in all of the evaluated metrics. These features may not be performing as well in our evaluation as was reported in the literature due to the limited size of the evaluated dataset. Here we are using videos 40 seconds long as a means of identifying FMs, whereas the supporting literature makes use of videos 3 to 5 minutes in duration to provide a final prediction of CP. Additionally, given that the assessed features are unable to model individual body parts, there may be some difficulty in the system identifying the specific movements which can be attributed to FMs. The optical flow and pose based methods appear to perform significantly better, with perfect classification performance (1.0000 Accuracy, 1.0000 Sensitivity, and 1.0000 Specificity) achieved by the Absolute Motion Distance (AMD), Relative Frequency (Freq.), Histogram of Joint Orientation (HOJO2D) and the fused Histogram of Joint Orientation and Joint Displacement (HOJO+JD). We also find that our method matches the best performing features, also reporting 100% in the accuracy, sensitivity and specificity metrics. This improved performance may be due to the ability of these methods to model the additional motion detail required to determine the presence of FMs in shorter sequences. We suggest that these more advanced computer vision techniques are better able to deal with localised motion patterns, rather than modelling the holistic movement based around a central mass.

When we interpret the results using the RVI-25 dataset we find a similar pattern to that of the MINI-RGBD dataset. We do however, also note that there is a general deterioration in classification performance suggesting that the RVI-25 dataset is more challenging; likely due to camera movement, illumination changes, noise and occlusion not found in the MINI-RGBD dataset. We observe that the Centroid of Motion and Quantity of Motion features again do not perform as well, with particularly low sensitivity throughout, despite using manually screened and cropped videos as input to the system. This is likely due to the aforementioned limitations as well as the additional challenges in the RVI-25 dataset. Given that the dataset was captured using a handheld camera to simulate smartphone recording 'in-the-wild', we see a negative impact in performance for these features, likely due to camera movements being incorrectly interpreted as infant motion. We find that the optical flow and pose-based methods are better able to model the infant motion, due to the localised nature of the analysis rather than

a full frame-based representation of the movement. Encouraging results are obtained from our method in all metrics, with the highest reported accuracy (0.9200) and sensitivity (0.8333). We also report the second highest specificity (0.9473), however the features which outperform our method in this metric offer poorer classification accuracy and perform significantly worse in the sensitivity metric. This increased performance is likely due to the deep learning architecture identifying complementary information, and subsequently capturing behaviors which are not fully detected through optical flow and pose estimation histogram representation. For example, we explicitly model the movements of each joint temporally using LSTM layers. In contrast, the previous methods use histogram-based representations which accumulate the values but discard the sequential information of body movement. This promising performance across all metrics suggests that our proposed system can not only provide state-of-the-art classification performance, but can also deal with the challenges inherent in the RVI-25 dataset whilst simultaneously allowing for our visualization module to function for interpretability and user feedback.

D. VISUALIZATION RESULTS AND DISCUSSION

We further provide qualitative results to demonstrate the effectiveness of our proposed framework. As presented in Section IV-D, we compute the attention values for each of the body part streams in order evaluate the contribution from each stream towards the overall classification (i.e. prediction in our case). The contribution values are then converted into the intensity of the highlighted text value (colored green for 'FM+' and red for 'FM−') of each body part. An example is illustrated in Figure 5. Readers are referred to the accompanying video demos to evaluate the visual quality of the results.

From the results, it can be seen that the highlighted body-parts generally show a lot of movement in the videos classified as negative (i.e. 'FM+' in Figure 5 (b)) or less complex, more repetitive movements in the videos classified as positive (i.e. 'FM−' in Figure 5 (a)). Specifically, in Figure 5(a), the upper body of the infant demonstrated a lack of FMs. As a result, the arms and upper legs are having larger contribution scores (i.e. higher intensity in red). This aligns well with our perception of GMA and demonstrates the effectiveness of our approach. The visualization provides effective visual feedback to the user, as such clinicians can pay greater attention to the highlighted segments for further analysis.

Our study has confirmed that by capitalizing on recent advances in deep learning we are able to successfully model an infant's movement patterns, quantifying CP risk-related FMs in variable clinical conditions. Our evaluation also suggests that with more data our system would be able to achieve even greater efficacy. Furthermore, our integrated approach to ensuring that our system is interpretable ensures that clinicians are kept at the forefront of this research. Whilst our system is able to provide rudimentary visual feedback to the user, additional visualization tools would be useful to exploit the extracted spatio-temporal information and provide additional predictive aid to clinicians for this complex diagnostic task. We suggest that by further integrating data augmentation techniques such as Synthetic Minority Oversampling (SMOTE) [13], Mixup [49] or temporal segmentation [34] we may be able to generate more data samples, further reducing overfitting and subsequently enhancing the classification and visualization performance of our proposed framework.

VI. CONCLUSION

In this paper, we have proposed a classification framework for the identification of fidgety movements associated with the prediction of cerebral palsy from video data. Our method takes advantage of the semantic context in the video, modelling body part-based movement using deep neural networks. To make the system more interpretable as a medical assessment tool, we also propose an automated visualisation module using motion the information extracted from individual body parts. Our visualization framework separates the input video frames into 8 body-part specific streams, making it possible to highlight which body part is contributing more toward the classification decision. As a result, our proposed framework provides users with additional diagnostic information, and helps with the interpretability problem common in machine learning approaches.

To facilitate evaluation of the proposed framework, we constructed a new and challenging dataset consisting of videos of pre-term infants deemed to be at higher risk of developing cerebral palsy. The videos were captured as part of routine clinical care, and were filmed from an overhead position with the infants lying in a supine position facing the camera. Experimental results show that our proposed method performs with comparable accuracy and greater robustness than the baseline methods, whilst additionally providing an interpretable visualization of the factors affecting classification.

Based upon our encouraging results we intend to further this work by gathering a larger, more extensively annotated dataset to extend our classification and visualization framework. This would minimise model overfitting problems and help consolidate our results. We hope to develop a tool with greater analytic feedback, and to enhance the proposed framework by incorporating additional features to further evaluate the spatio-temporal relationships found in movement disorders. By enhancing the temporal aspect of our classification and visualization framework, we hope to be able to highlight temporal time-frames of clinical interest, to draw attention to moments within a sequence where risk related movements are present.

Moreover, existing research formulates the CP prediction problem as a binary classification in which the data samples are annotated with a single label (i.e. positive or negative). Alternatively, annotating data at a finer level, for example, specifying the presence or absence of fidgety movement at each body part over time can certainly provide useful information to improve the performance of our proposed

framework. In the future, we are interested in inviting clinicians to provide us with annotations at such a level. Approaches such as *active learning* will be explored to obtain annotation from the clinicians strategically to minimize the manual work. Another interesting future direction will be exploring weakly supervised learning approaches to tackle the aforementioned data annotation problem.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] L. Adde, J. Helbostad, A. R. Jensenius, M. Langaas, and R. Støen, "Identification of fidgety movements and prediction of CP by the use of computer-based video analysis is more accurate when based on two video recordings," *Physiotherapy Theory Pract.*, vol. 29, no. 6, pp. 469–475, Aug. 2013.

[2] L. Adde, J. L. Helbostad, A. R. Jensenius, G. Taraldsen, K. H. Grunewaldt, and R. Støen, "Early prediction of cerebral palsy by computer-based video analysis of general movements: A feasibility study," *Develop. Med. Child Neurol.*, vol. 52, no. 8, pp. 773–778, Feb. 2010.

[3] L. Adde, M. Rygg, K. Lossius, G. K. Øberg, and R. Støen, "General movement assessment: Predicting cerebral palsy in clinical practise," *Early Hum. Develop.*, vol. 83, no. 1, pp. 13–18, Jan. 2007.

[4] L. Adde, H. Yang, R. Sæther, A. R. Jensenius, E. Ihlen, J.-Y. Cao, and R. Støen, "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.

[5] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, Oct. 2014.

[6] D. Ahmedt-Aristizabal, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, and C. Fookes, "Understanding patients' behavior: Vision-based analysis of seizure disorders," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2583–2591, Nov. 2019.

[7] 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.

[8] *What is the General Movements Assessment*, Cerebral Palsy Alliance, Allambie Heights, NSW, Australia, 2018.

[9] M. Bax, M. Goldstein, P. Rosenbaum, A. Leviton, N. Paneth, B. Dan, B. Jacobsson, and D. Damiano, "Proposed definition and classification of cerebral palsy, April 2005," *Develop. Med. Child Neurol.*, vol. 47, no. 8, pp. 571–576, 2005.

[10] 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.

[11] 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.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. CVPR*, Jul. 2017, pp. 7291–7299.

[13] V. N. Chawla, W. K. Bowyer, O. L. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique Nitesh," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Sep. 2002.

[14] C. Einspieler, A. F. Bos, M. E. Libertus, and P. B. Marschik, "The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction," *Frontiers Psychol.*, vol. 7, Mar. 2016. [Online]. Available: https://doi.org/10.3389%2Ffpsyg.2016.00406, doi: 10.3389/fpsyg.2016.00406.

[15] C. Einspieler and H. F. R. Prechtl, "Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system," *Mental Retardation Developmental Disabilities Res. Rev.*, vol. 11, no. 1, pp. 61–67, 2005, doi: 10.1002/mrdd.20051.

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

[17] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 280–291, Jan. 2020.

[18] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[19] T. Fjørtoft, C. Einspieler, L. Adde, and L. I. Strand, "Inter-observer reliability of the 'Assessment of motor repertoire-3 to 5 months' based on video recordings of infants," *Early Human Develop.*, vol. 85, no. 5, pp. 297–302, 2009.

[20] *NICE Seeks to Improve Diagnosis and Treatment of Cerebral Palsy*, Nat. Inst. Health Care Excellence, London, U.K., Jan. 2017.

[21] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[22] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[23] 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 *Computer Vision—ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham, Switzerland: Springer, 2019, pp. 32–49, doi: 10.1007/978-3-030-11024-6_3.

[24] A. Holzinger, C. Biemann, S. C. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*. [Online]. Available: https://arxiv.org/abs/1712.09923

[25] E. A. F. Ihlen, R. Støen, L. Boswell, R.-A. de Regnier, T. Fjørtoft, D. Gaebler-Spira, C. Labori, M. C. Loennecken, M. E. Msall, U. I. Möinichen, C. Peyton, M. D. Schreiber, I. E. Silberg, N. T. Songstad, R. T. Vågen, G. K. Øberg, and L. Adde, "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.

[26] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, 2021, Art. no. 107398. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020307846, doi: 10.1016/j.ymssp.2020.107398.

[27] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," Tech. Rep., 2017.

[28] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 314–323, Jan. 2019.

[29] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5137–5146.

[30] J. Manttari, S. Broome, J. Folkesson, and H. Kjellstrom, "Interpreting video features: A comparison of 3D convolutional networks and convolutional LSTM networks," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020.

[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] 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.

[33] 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.

[34] K. McCay, E. S. L. Ho, D. Sakkos, W. L. Woo, C. Marcroft, P. Dulson, and N. Embleton, "Towards explainable abnormal infant movements identification: A body-part based prediction and visualisation framework," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI) (IEEE BHI)*, Athens, Greece, Jul. 2021.

[35] C. Morgan, C. Crowle, T.-A. Goyen, C. Hardman, M. Jackman, I. Novak, and N. Badawi, "Sensitivity and specificity of general movements assessment for diagnostic accuracy of detecting cerebral palsy early in an Australian context," *J. Paediatrics Child Health*, vol. 52, no. 1, pp. 54–59, Jan. 2016.

[36] S. Orlandi, K. Raghuram, C. R. Smith, D. Mansueto, P. Church, V. Shah, M. Luther, and T. Chau, "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.

[37] 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.
[38] P. C. Panteliadis, *Cerebral Palsy: A Multidisciplinary Approach*, 3rd ed. Springer, 2018.
[39] H. F. Prechtl, C. Einspieler, G. Cioni, A. F. Bos, F. Ferrari, and D. Sontheimer, "An early marker for neurological deficits after perinatal brain lesions," *Lancet*, vol. 349, no. 9062, pp. 1361–1363, May 1997.
[40] 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.
[41] S. M. Reid, E. Meehan, S. McIntyre, S. Goldsmith, N. Badawi, D. S. Reddihough, and The Australian Cerebral Palsy Register Group, "Temporal trends in cerebral palsy by impairment severity and birth gestation," *Develop. Med. Child Neurol.*, vol. 58, no. S2, pp. 25–35, 2016.
[42] P. Rosenbaum, N. Paneth, A. Leviton, M. Goldstein, M. Bax, D. Damiano, B. Dan, and B. Jacobsson, "A report: The definition and classification of cerebral palsy April 2006," *Develop. Med. Child Neurol. Suppl.*, vol. 109, pp. 8–14, Mar. 2007.
[43] A. Stahl, C. Schellewald, Ø. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerod, "An optical flow-based method to predict infantile cerebral palsy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 4, pp. 605–614, Jul. 2012.
[44] S. Suwajanakorn, N. Snavely, J. Tompson, and M. Norouzi, "Discovery of latent 3D keypoints via end-to-end geometric reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, no. 1, 2018, pp. 2059–2070.
[45] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4627–4635.
[46] 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.
[47] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing R-CNN for instance-level human analysis," in *Proc. CVPR*, 2019, pp. 364–373.
[48] L. Yao, Y. Liu, and S. Huang, "Spatio-temporal information for human action recognition," *EURASIP J. Image Video Process.*, vol. 2016, no. 1, p. 39, Nov. 2016.
[49] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb and https://dblp.org/rec/conf/iclr/ZhangCDL18.bib

DIMITRIOS SAKKOS received the B.Sc. degree in mathematics from the Aristotle University of Thessaloniki, Greece, in 2012, the M.Sc. degree in computer science from the University of Birmingham, U.K., in 2013, and the Ph.D. degree in computer science from Northumbria University, in 2020.

His research interest includes in the field of image and video segmentation.

KEVIN D. MCCAY received the B.Sc. degree (Hons.) in computer animation and digital SFX and the M.A. degree in animation from Northumbria University, Newcastle upon Tyne, U.K., in 2015 and 2016, respectively, where he is currently pursuing the Ph.D. degree in computer science.

His research interests include human motion analysis, computer vision, machine learning, and the application of automated analysis within the healthcare domain.

CLAIRE MARCROFT received the B.Sc. degree (Hons.) in physiotherapy from the University of Huddersfield, Huddersfield, U.K., in 2004.

She is a Neonatal Physiotherapist at Newcastle upon Tyne Hospitals, NHS Foundation Trust/Newcastle University.

Ms. Marcroft is a member of the Health Care and Professions Council (HCPC), the Chartered Society of Physiotherapy (CSP), and the Association of Paediatric Chartered Physiotherapists (APCP). She holds a National Institute of Health Research (NIHR) Integrated Clinical Academic Doctoral Research Fellowship (ICA-CDRF-2018-04-ST2-020).

NICHOLAS D. EMBLETON received the B.Sc. degree in environmental studies and computer science from the University of East Anglia, Norwich, Norfolk, U.K., in 1984, and the Medicine M.B.B.S. degree (Hons.) and the Doctor of Medicine M.D. degree (Hons.) from Newcastle University, Newcastle upon Tyne, U.K., in 1990 and 2002, respectively.

He is a consultant neonatal paediatrician and a professor of neonatal medicine. He completed paediatric and neonatal training in U.K., and Vancouver, Canada. He helps to lead a broad portfolio of research coordinated by the Newcastle Neonatal Research Team which includes large-scale NIHR nutrition trials, along with mechanistic microbiomic and metabolomic studies. He coordinates the Newcastle Preterm Birth Growth Study that has tracked the growth and metabolic outcomes of children who were born preterm into late adolescence. He leads a series of qualitative studies exploring the experiences of parents who suffered a reproductive or neonatal loss. He has more than 140 peer reviewed publications in addition to numerous educational articles and book chapters.

Dr. Embleton is an elected member of the ESPGHAN Committee of Nutrition and coordinates the U.K.-based Neonatal Nutrition Network.

SAMIRAN CHATTOPADHYAY (Senior Member, IEEE) received the bachelor's and master's degrees in computer science and engineering from IIT Kharagpur, India, and the Ph.D. degree from Jadavpur University, Kolkata, India. He is currently working as a Professor with the Department of Information Technology, Jadavpur University. He is having over 25 years of teaching experience at Jadavpur University, four years of industry experience, and 12 years of technical consultancy in the reputed industry houses. He has authored over 110 articles in international journals and papers in conferences.

EDMOND S. L. HO received the B.Sc. degree in computer science from Hong Kong Baptist University, the M.Phil. degree from the City University of Hong Kong, and the Ph.D. degree from the University of Edinburgh.

He is currently a Senior Lecturer with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, U.K. Prior to joining Northumbria University, in 2016, he was a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University. His research interests include computer graphics, computer vision, robotics, motion analysis, and machine learning.