Cerebral Palsy Prediction with Frequency Attention Informed Graph Convolutional Networks

Haozheng Zhang[1], Hubert P. H. Shum[2] and Edmond S. L. Ho[3]

***Abstract*— Early diagnosis and intervention are clinically considered the paramount part of treating cerebral palsy (CP), so it is essential to design an efficient and interpretable automatic prediction system for CP. We highlight a significant difference between CP infants' frequency of human movement and that of the healthy group, which improves prediction performance. However, the existing deep learning-based methods did not use the frequency information of infants' movement for CP prediction. This paper proposes a frequency attention informed graph convolutional network and validates it on two consumer-grade RGB video datasets, namely MINI-RGBD and RVI-38 datasets. Our proposed frequency attention module aids in improving both classification performance and system interpretability. In addition, we design a frequency-binning method that retains the critical frequency of the human joint position data while filtering the noise. Our prediction performance achieves state-of-the-art research on both datasets. Our work demonstrates the effectiveness of frequency information in supporting the prediction of CP non-intrusively and provides a way for supporting the early diagnosis of CP in the resource-limited regions where the clinical resources are not abundant.**

I. INTRODUCTION

General Movement Assessment (GMA) [1] is being widely used clinically for the early prediction of cerebral palsy (CP). However, targeted GMA training for clinicians is a time-consuming and resource-consuming task. As a result, only a small but increasing number of clinicians have received this training in the UK and Australia [2]. Furthermore, the process also requires manual inspection of the infant movement and is prone to subjective assessment. Early studies applied machine learning techniques (e.g. support vector machine, random forest) and the optical flow-based video analysis method to propose the automated GMA systems [3], [4]. But these works still require manual labelling of infant joint positions. Some later studies focus on the analysis of frequency domain data. Stahl *et al*. [5] used an optical flow-based approach to assess infant movements and then applied wavelet frequency analysis to evaluate the time-dependent trajectory signals in optical flow data. Rahmati *et al*. [6] applied a motion segmentation algorithm to extract motion data from each limb in the infant video and then classified the infants' movements with features obtained by frequency analysis.

*This work was supported in part by the Royal Society (Ref: IES\R2\181024 and IES\R1\191147)

[1]Haozheng Zhang is with Durham University, the United Kingdom haozheng.zhang@durham.ac.uk

[2]Hubert P. H. Shum is with Durham University, the United Kingdom. Corresponding author. hubert.shum@durham.ac.uk

[3]Edmond S. L. Ho is with Northumbria University, the United Kingdom e.ho@northumbria.ac.uk

Recent deep learning-based systems achieved impressive performance in CP infants movement prediction. McCay *et al.* [7] proposed a fully connected deep learning network and four Convolutional Neural Network (CNN)-based deep learning architectures to classify the abnormal movements of CP infants by using the histogram of joint orientation 2D and joint displacement 2D features, achieved the highest prediction accuracy of 91.67% on the MINI-RGBD dataset [8]. Zhu [9] further applied the channel attention mechanism on the 2D-CNN model to interpret the CP prediction outcome on the same dataset. However, the robustness and generality of their proposed method have not been fully evaluated since the results are obtained from a single small dataset.

Aiming at the significant difference in joints movement frequency between the cerebral palsy infants and the healthy group, in this article, we demonstrate a frequency-based binning mechanism and a graph convolution network to improve the performance of CP prediction with better interpretability. Firstly, we employ a pose estimation algorithm, namely Openpose [10] to extract the human joint position data from the R-GBD video sequences as the input to our system. Then we propose an automatic frequency-binning module suitable for videos with different frame rates to reduce data noise and the percentages of high-frequency movements information in the whole video sequence for CP prediction. The idea is inspired by both the frequency analysis-based infants CP prediction methods [6] and our observation. Rahmati *et al*. [6] provided a result that comparing with very low or high-frequency ranges, the middle-to-low frequency range data showed more differences between the healthy group and the CP group. In addition, we found that the infants' joint position data in the high-frequency domain is mainly caused by data noise, such as the misdetected joint position by Openpose.

We validate our system on the MINI-RGBD dataset [11] and the RVI-38 dataset [12]. The MINI-RGBD dataset has been widely used for CP classification performance comparison in the previous work [7], [13], [14], [15], including synthetic video sequences of 12 normal and CP infants. The RVI-38 is a recently collected dataset for a more challenging CP prediction task, with a larger size of data captured during routine clinical care. Experimental results show that our system achieves state-of-the-art CP prediction performance on both of the dataset and allows users to interpret the weights of movement frequencies of different joints in our prediction system.

Our contributions are as follows:

- We interpret the Cerebral Palsy prediction in the joint

movement frequency domain by the attention module. In addition, we designed a new frequency-binning module that can be applied to both deep learning and machine learning networks for videos with different frame rates to improve the CP prediction performance.
- We propose a novel frequency attention informed graph convolutional network (FAIGCN) for CP prediction from consumer-grade RGB-D videos. Our system achieves state-of-the-art research on two datasets with strong robustness.
- We open our source code for validation and further development: https://github.com/zhz95/FAIGCN

II. Dataset Processing

We verify our models on the Moving INfants In RGB-D synthetic dataset (MINI-RGBD) [11] and RVI-38 dataset.

A. The MINI-RGBD Dataset

MINI-RGBD was generated by registering and rendering the synthetic Skinned Multi-Infant (SMIL) model [8] to the RGB-D sequences of real-world moving infants recorded in the hospital. All 12 RGB-D video sequences were captured when the infants were half-year-old. The MINI-RGBD dataset is a popular open resource relating to infants CP as it consists of realistic shape, texture and movement. It also provides precise ground truth while anonymizing the data by replacing the raw video frames with computer graphics rendered frames. We further obtained the annotation of each video sequence shared by [13], which indicates the presence (i.e. labelled as "normal") or absence (i.e. labelled as "abnormal") of fidgety movements in the video by an independent medical expert using the GMA method [1].

B. The RVI-38 Dataset

The RVI-38 dataset was collected from a part of routine clinical care at the Royal Victoria Infirmary (RVI) in Newcastle upon Tyne, UK. There are 38 RGB-D video sequences of different infants between 36-60 weeks in the RVI-38 dataset. All videos were captured by a consumer-grade handheld camera (Sony DSC-RX100 with a resolution of 1980x1080 and the 25FPS frame rate). The length of videos ranges between 40 seconds and 5 minutes, with an average length of 3 minutes and 36 seconds. The camera was set above the baby, and the infant's movement was photographed from top to bottom. All videos were annotated using the GMA method by two experienced assessors. The annotations indicate the presence (i.e. labelled as "normal") or absence (i.e. labelled as "abnormal") of fidgety movements in the video.

C. Data Preprocessing

For more effective CP predictions, we extract 2D skeleton features from the video sequences. We apply OpenPose [10] for pose estimation due to its high accuracy in detecting the posture of the infants, and it is less sensitive to variations on the appearance. OpenPose returns the 2D coordinates (x, y) for 18 human joint landmarks and a confidence score C for each joint estimation. However, for joints that are self-occluded or without clear visual features, OpenPose would not be able to deduce their position, and zero values would be returned as the joint positions. As this may impact the performance of the prediction system, we propose to preprocess the data by replacing the zero values in frame f with the linear interpolation of neighbouring non-zero frames.

In order to overcome the overfitting in small size dataset, we implement several processes. 1) We calculate the global normalization of joint positions frame by frame to reduces the infant's global translation. To achieve this, we set the center of the triangle of the neck and two hip joints as the global origin, then relocate each joints by the relative distance between joints. 2) To normalize the x-direction and y-direction pose features, we align the line between the global orgin and the neck joint with the y-axis and keep the neck joint above the global origin.

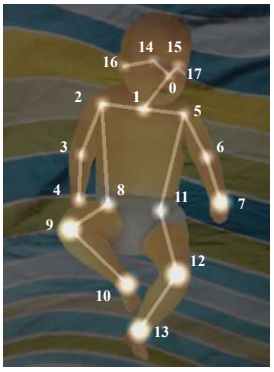

Fig. 2. An example frame of 18-joints Openpose [10] posture layout for infant kicking in the MINI-RGBD dataset [8]. The size of the light point represents the size of the attention value, that is, the importance of the movement frequency of the joint to our network in CP prediction at that frame.

III. Frequency Attention Informed GCN System

The proposed system consists of two parts (seen in Fig.1): (1) The frequency-binning module transforms the input joint movement features into the frequency domain, then filters high-frequency information to make our prediction network focus on low-to-mid infants movement frequencies. (2) The proposed Frequency Attention Informed GCN for CP prediction and interpretation. Fig. 2 shows an example of attention visualisation.

A. The Frequency-binning Module

Given the infants' joint movement sequences as input, we propose using frequency operations on joint position data for CP prediction. It is motivated by two observations. Firstly, the body movement frequencies of healthy infants are different from infants who suffered from CP [6], and the low-frequency range information of body movement is more critical for fidgety movements (FMs). FMs are moderate speed movement of the neck, trunk and limbs with different accelerations in various directions [16]. Previous work [16], [17] has shown that the absence of FMs is an essential distinguishing feature of CP infants from healthy infants.

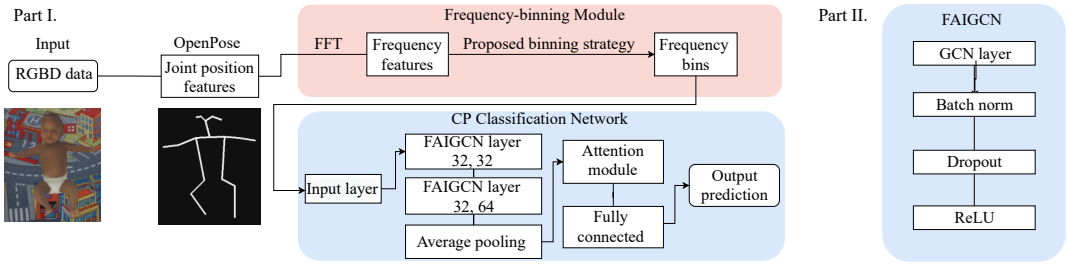

Fig. 1. The overview of our proposed framework. Part I is the overall network architecture, Part II is the design of each FAIGCN layer.

Frequency-binning can filter the high-frequency (e.g., above 6 HZ) information of joint position data after FFT, thus making the classification network focused on low-to-mid (e.g., 0-5 HZ) frequency infant movements without eliminating raw data. Secondly, the infant movements frequencies are generally low, and the high-frequency range from the joint position data is mainly due to data noise, such as the misdetected joint position and the video capture error from the datasets.

As a solution, we design a frequency-binning module that retains the critical frequency of the joint position data while filtering the noise. The module employs Fast Fourier Transform (FFT) to convert the time series of joint positions into the frequency domain, then applies frequency binning to obtain the motion frequency information mainly distributed in the low-to-mid band. This module is adaptable for videos with a frame rate between 24 FPS to 60 FPS and is suitable for both DNN or machine learning-based classification models. The core of the module is the binning strategy, in which we design a formula to use finer bins for the more crucial low-to-mid frequency and coaster bins for higher frequency.

1) Fast Fourier Transform (FFT): We apply Bluestein's FFT algorithm [18], a discrete Fourier transform algorithm, on all 2D joints movements time series to transform original joints position features into the frequency domain and obtain the frequency components:

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-i2\pi kn}{N}}, \quad k = 0, \ldots, N-1 \tag{1}$$

where x_n is a time series, $e^{\frac{i2\pi}{N}}$ is a primitive N^{th} root of 1.

2) The Binning Strategy: We propose a data binning strategy to emphasize the importance of low-frequency information of the joint position data. Under the strategy, the width of the bins are different - smaller width bins are used for low-frequency range and increasingly larger-width bins for higher frequency range:

$$b_n = \begin{cases} \text{Round}(b_0 \cdot c^n), & \text{if } b_n \cdot c^n < 3, \\ \text{Ceiling}(b_0 \cdot c^n), & \text{if } b_n \cdot c^n \geq 3, \end{cases} \tag{2}$$

where b_n is the width of the n^{th} bin, $b_0 = 1$, and c is a controllable parameter. Note that the width of each bin needs to be an integer as FFT is a discrete (i.e. integer-based) system. This equation takes the round of the bin width when the width is less than three units to increase the density of the bins in low-to-mid frequency. According to the characteristics of rapid exponential growth, this function distinguishes the density of the middle frequency band and the high-frequency band for bins with a width greater than two units by rounding up the value greater than three units. Empirically, as shown in Sec IV, we achieve the best prediction accuracy when $c = 1.00264$ for the 25 FPS videos. The parameter c could be automatically generated for the best binning results by achieving the highest CP prediction performance for different datasets.

As a result, after being processed by the frequency-binning module, input joint position data are transformed into the frequency domain and endowed with an important characteristic: low-to-mid frequency information occupies a significantly more prominent emphasis.

B. The CP Prediction Network

As shown in Fig.1 , we propose a Frequency Attention Informed Graph Convolutional Network (FAIGCN) for CP prediction by classifying low-to-mid frequency band infant movement frequency features with the attention mechanism.

1) Frequency Attention Informed Network: Most of the previous DNN-based studies on infants CP prediction are based on traditional Convolutional Neural Networks (CNN). However, traditional discrete convolution from CNN can only maintain translational invariance on Euclidean data, which is not suitable for graph structure data such as the human skeletal graph generated from OpenPose [10].

Therefore, we employ a GCN [19] to learn the infant joints dependencies from the pose graph. Inspired by [20], we apply the pose graph which align with the human skeletal graph $G = (V, E)$ for interpreting which joint's movement frequency features are considered to be important in CP prediction task. In this graph, $\{V = v_{bi} | b = 1, ..., B; i = 1, ..., N\}$ denotes the frequencies of all joints, where v_{bi} represents the b-th frequency bin of i-th joint. The edge set E includes: (1) the intra-skeleton connection at each frequency band, $\{v_{bi}v_{bj} | (i, j) \in K\}$, where K is designed by the natural connections of human joints. (2) the inter-frequency edges which connect the frequency bins of a joint in the low-to-high frequency order, $\{v_{bi}v_{(b+1)i}\}$.

The graph convolutional operation of FAIGCN is followed by [19], where the propagation rule between layers can be represented by Eq. 3.

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{A}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{3}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_L$ is known as the adjacency matrix of an undirected graph. $\mathbf{I}_L$ is an L dimensions identity matrix. $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$ and $\mathbf{W}^{(l)}$ is a learnable weight matrix specified to the layer. The nonlinear activation function $\sigma(\cdot)$ is set as *ReLU* in our network.

We propose the following frequency attention-informed mechanism to learn the weight of frequency features. We aggregate the frequency features obtained from the frequency-binning module $\{\mathbf{h}_{1,i}, \mathbf{h}_{2,i}, \ldots, \mathbf{h}_{B,i}\}$ with attentions $\alpha_{b,i}$ by Eq. 4

$$\mathbf{v}_k = \sum_{b=1}^{B} \alpha_{b,i} \mathbf{h}_{b,i} \tag{4}$$

in which the frequency attention weight $\alpha_{b,i}$ is defined as:

$$\alpha_{b,i} = \frac{\exp\left(\sigma_n^{'}\left(\mathbf{w}_\alpha^{\mathsf{T}}, \mathbf{z}_{b,i}\right)\right)}{\sum_b \exp\left(\sigma^{'}\left(\mathbf{w}_\alpha^{\mathsf{T}}, \mathbf{z}_{b,i}\right)\right)} \tag{5}$$

$$\mathbf{z}_{b,i} = tanh\left(\mathbf{W}_z \mathbf{h}_{b,i}\right) \tag{6}$$

where $\sigma_n^{'}$ is an adjustable activation function as follows:

$$\sigma_n^{'} = \begin{cases} 1 + \left(\frac{\mathbf{w}_\alpha}{\|\mathbf{w}_\alpha\|}\right)^{\mathsf{T}} \left(\frac{\mathbf{z}_\alpha}{\|\mathbf{z}_\alpha\|}\right) & , n = 1 \\ \mathbf{w}_\alpha^{\mathsf{T}} \mathbf{z}_\alpha & , n = 2 \end{cases} \tag{7}$$

where $\mathbf{w}_\alpha$ and $\mathbf{W}_z$ are learnable parameters.

2) Network Adaptation: As can be seen from Fig.1, the input layer transforms the tensor format of input frequency data to fit in the network. Then, we use two FAIGCN layers with 32, 64 output channels respectively. Each FAIGCN, in turn, consists of a GCN layer, a batch normalization layer, a dropout and a ReLU layer. The kernel sizes of FAIGCN layers $K = 3, 3$, and $stride = 1, 2$, respectively. We put a global pooling layer after two FAIGCN layers. We applied the average pooling as it provides the highest robustness. At the last, we put a fully connected layer to classify features for CP prediction. The optimizer is chosen as *Adams*, and we train the model with *batch size* $= 1$, *learning rate* $= 0.0001$ with 0.1 decay every 100 epoches, *Max Epoch* $= 500$ on the MINI-RGBD dataset; *batch size* $= 4$, *learning rate* $= 0.001$ with 0.1 decay every 100 epoches, *Max Epoch* $= 500$ on the RVI-38 dataset.

IV. Experiments

Our experiments were run on a PC with Ubuntu 18.04 and an NVIDIA GeForce RTX 3080. The total model training time on MINI-RGBD with 12 sequences is about an hour, including estimation of the joints position from RGB videos. But it only takes about 50s for the CP prediction of 1000 frames ($\sim$ 33s) video sequence, which can be employed in interactive-time diagnosis.

A. Experimental Settings

In this paper, we conduct the leave-one-out cross-validation among two datasets to evaluate our proposed system. This setting utilises all data and ensures that the prediction system is evaluated against unseen data. Our evaluation metrics are introduced in Sec IV-C. We report the best result for each method to be consistent with several related works in literature [5], [17], [6], [15], [12].

B. Comparing with State-of-the-art Methods

In order to evaluate the effectiveness of our system, we compare FAIGCN with the following methods:

- **FCNet** [7]: This method uses fully connected deep network architectures to the Histogram of Joint Displacement 2D (HOJD2D) and Histogram of Joint Orientation 2D (HOJO2D) calculated from human joint positions. For the HOJD2D feature, the displacements of each joint are extracted every five frames and segmented into 16 bins. The feature of HOJO2D represents the joint orientation in 2D space, and the feature is also segmented into 16 bins.
- **Conv1D-1**, **Conv1D-2** [7]: They are two 1D convolutional neural networks, each of them consists of two 1D convolutional layers with differences in the output channel sizes. They are proposed to classify the abnormal infant movements by feature HOJO2D or HOJD2D (HOJO/D2D).
- **Conv2D-1**, **Conv2D-2** [7]: They are two 2D convolutional neural networks, each of them consists of two 2D convolutional layers with differences in the output channel sizes.
- **CANet** [9]: This method proposes a 2D convolutional neural network with the squeeze-and-excitation channel attention module. The whole system is proposed for the CP classification task on the MINI-RGBD dataset.
- **ST-GCN** (Spatial Temporal Graph Convolutional Network) [20]: This is a graph convolutional neural network for human skeleton data (e.g. joint position).
- **STAM** [21]: This is a spatial-temporal graph convolutional neural network with the attention mechanism.
- **Ens-1** [13]: This method uses an ensemble classifier on the fused feature of HOJO2D + HOJD2D (HOJO+D2D) with eight bins.
- **Ens-2** [12]: This method extends [13] by fusing four pose-related features and three velocity-related features.
- **Ens-3** [22]: This method extends [13] by extracting features at limb-level from small video segments to locate abnormal movements spatiotemporally.
- **MCI** [14]: This method uses a threshold model to classify the infant CP via Movement Complexity Index (MCI), where MCI is computed by extracting the infant's limb angle features.

C. Evaluation Metrics

We evaluate our system and other state-of-the-art methods by following five metrics: the prediction accuracy (AC) shows the percentage of correctly predicted individuals in the

dataset; the sensitivity (SE) shows the percentage of correctly predicted positive individuals among the total number of positive individuals in the dataset; the specificity (SP) shows the percentage of correctly predicted negative individuals among the total number of negative individuals in the dataset; F1-Score evaluates the binary classification performance by calculating the harmonic mean of the precision and recall; Matthews Correlation Coefficient (MCC) [23] provides a reliable performance metric for imbalanced dataset [24].

TABLE I

THE COMPARISON WITH STATE-OF-THE-ARTS ON THE MINI-RGBD

Method	Feature	AC	SE	SP	F1	MCC
FCNet [7]	HOJD2D	91.67	**100.00**	87.50	88.89	83.67
FCNet [7]	HOJO2D	83.33	75.00	87.50	75.00	62.50
Conv1D [7]	HOJO/D2D	83.33	75.00	87.50	75.00	62.50
Conv2D [7]	HOJO/D2D	83.33	75.00	87.50	75.00	62.50
Conv2D [7]	HOJO+D2D	91.67	**100.00**	87.50	88.89	83.67
Conv2D [7]	Pose	83.33	75.00	87.50	75.00	62.50
CANet [9]	Pose	91.67	**100.00**	87.50	88.89	83.67
ST-GCN[20]	Pose	91.67	**100.00**	87.50	88.89	83.67
STAM [21]	Pose	91.67	**100.00**	87.50	88.89	83.67
Ens-1 [13]	HOJO+D2D	91.67	**100.00**	87.50	88.89	83.67
Ens-2 [12]	Velocity	91.67	**100.00**	87.50	88.89	83.67
Ens-2 [12]	Pose*	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Ens-2 [12]	Vel.+Pose*	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Ens-3 [22]	HOJO+D2D	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
MCI [14]	Limb angle	91.67	**100.00**	87.50	88.989	83.67
FAIGCN	Motion Freq.	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

* The Pose and velocity features here fuses several hand-crafted features including HOJOD2D

D. Result Comparison

We report the prediction results on MINI-RGBD and RVI-38 datasets on Table I and Table II respectively. From our evaluation, we propose the following observations: (1) Our FAIGCN system outperforms the state-of-the-art DNN based methods in both datasets. Comparing with other non-DNN based methods, our system also achieves state-of-the-art performance in two datasets. (2) We can observe the advantage of the attention mechanism in DNN as CANet outperforms Conv2D-Pose from Table I and Table II, and CANet outperforms Conv1D/Conv2D, STAM outperforms ST-GCN from Table II. (3) From Table II, it can be seen that ST-GCN, STAM and FAIGCN outperform all CNN-based methods (i.e. Conv1D, Conv2D and CANet), which confirms the advantage of using graph structure to analyse human pose data. (4) We notice that the methods that use early fusion on features outperform those using only a single kind of feature. It can be seen by comparing Conv2D-HOJO/D2D with Conv2D-HOJO+D2D, or comparing Ens-2-Velocity/Pose with Ens-2-Velocity+Pose in both tables. Therefore, we consider fusing our movement frequency features with other features in future work. (5) An interesting finding is that the Machine learning-based methods such as Ens-1, Ens-2 and Ens-3 outperform DNN-based methods except for FAIGCN. On the one hand, it shows the superiority of hand-craft features in the classification tasks; On the other hand, it inspires us to compare FAIGCN with machine learning-based methods with the same features, seen in the Sec. IV-E below.

TABLE II

THE COMPARISON WITH STATE-OF-THE-ARTS ON THE RVI-38

Method	Feature	AC	SE	SP	F1	MCC
Conv2D [7]	Pose	81.58	33.33	90.63	36.36	25.85
CANet [9]	Pose	86.84	66.67	90.63	61.54	53.89
ST-GCN [20]	Pose	89.47	66.67	93.75	62.50	60.42
STAM [21]	Pose	92.11	**83.33**	93.75	76.92	72.51
Ens-1 [13]	HOJO+D2D	94.74	**83.33**	96.88	83.33	80.21
Ens-2 [12]	Velocity	94.74	**83.33**	96.88	83.33	80.21
Ens-2 [12]	Pose*	94.74	**83.33**	96.88	83.33	80.21
Ens-2 [12]	Vel.+Pose*	**97.37**	**83.33**	**100.00**	**90.91**	**89.89**
FAIGCN	Motion Freq.	**97.37**	**83.33**	**100.00**	**90.91**	**89.89**

* The Pose and velocity features here fuses several hand-crafted features including HOJOD2D

E. Ablation study

The ablation study is divided into two parts:

In the first part, we evaluate whether there is any adverse effect on prediction performance caused by the frequency-binning module (B.) or the attention module (A.). The corresponding result is displayed in Table III. Due to the limitation of the MINI-RGND dataset's size, the contribution of the attention mechanism is not reflected significantly. But from the larger RVI-38 dataset, it can be seen that there is a significant improvement by applying the attention module or frequency binning module.

In the second part, we implement four machine learning-based methods with a proposed frequency-binning module to predict the CP using the movement frequency features from both datasets. The methods are Support Vector Machine (SVM), Decision Tree (Tree), Logistic Regression (LR) and Linear Discriminant Analysis (LDA). The ensemble of classification models Ens-1 [13], Ens-2 [12] and Ens-3 [22] are not included since the types of the ensemble classifier in Matlab was used which consists of a wide range of classifiers and handles the late-fusion internally. Besides, we validate the effectiveness and robustness of the frequency-binning module by eliminating it from each method. The results are reported in Table IV. We observe the advantage of using the proposed frequency-binning module as each method outperforms its variant of no frequency-binning module, except in the case of SVM in the MINI-RGBD dataset. In addition, we note that our system outperforms the implemented machine learning-based methods. It demonstrates the effectiveness of graph convolutional neural network in dealing with the same features.

TABLE III

THE PERFORMANCE OF FAIGCN AND ITS SIMPLIFIED VARIANTS

The MINI-RGBD dataset					
Method	AC	SE	SP	F1	MCC
FAIGCN-full	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
w/o A.	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
w/o B.	91.67	**100.00**	87.50	88.89	83.67
w/o A. B.	91.67	**100.00**	87.50	88.89	83.67

The RVI-38 dataset					
Method	AC	SE	SP	F1	MCC
FAIGCN-full	**97.37**	**83.33**	**100.00**	**90.91**	**89.89**
w/o A.	92.11	**83.33**	93.75	76.92	72.51
w/o B.	89.47	66.67	93.75	62.50	60.42
w/o A. B.	86.84	66.67	90.63	61.54	53.89

TABLE IV

THE COMPARISON WITH MACHINE LEARNING BASED METHODS AND THEIR VARIANT WITHOUT FREQUENCY-BINNING MODULE

The MINI-RGBD dataset					
Methods	AC	SE	SP	F1	MCC
SVM	66.77	75.00	62.50	66.67	35.36
SVM w/o B.	66.77	75.00	62.50	66.67	35.36
Tree	75.00	75.00	75.00	66.67	47.81
Tree w/o B.	66.77	75.00	62.50	66.67	35.36
LDA	83.33	75.00	87.50	75.00	62.50
LDA w/o B.	75.00	75.00	75.00	66.67	47.81
LR	91.67	100.00	87.50	88.89	83.67
LR w/o B.	75.00	75.00	75.00	66.67	47.81
FAIGCN-full	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
FAIGCN w/o B.	91.67	**100.00**	87.50	88.89	83.67

The RVI-38 dataset					
Methods	AC	SE	SP	F1	MCC
SVM	63.16	66.67	62.50	66.67	35.36
SVM w/o B.	55.26	50.00	56.25	26.09	4.58
Tree	81.58	66.67	84.38	53.33	43.78
Tree w/o B.	68.42	50.00	71.89	33.33	17.16
LDA	83.33	75.00	87.50	75.00	62.50
LDA w/o B.	65.79	66.67	65.63	38.10	24.09
LR	78.95	66.67	81.25	50.00	39.68
LR w/o B.	57.89	66.67	56.25	33.33	16.74
FAIGCN-full	**97.37**	**83.33**	**100.00**	**90.91**	**89.89**
FAIGCN w/o B.	89.47	66.67	93.75	62.50	60.42

F. Robustness Test

In order to evaluate the robustness of our system and other state-of-the-art DNN-based methods [9], [20], [21], we simulate different datasets by adding different levels of Gaussian noise to the infant joint pose data. The noise level is divided into four levels from 15% standard deviation to 120% standard deviation of each infant's joint pose data. The tests results are displayed in Fig. 3. The accuracy in the y-axis is the average accuracy among ten leave-one-out cross-validations with ten different training seeds.

From Fig. 3, we observe that as the noise level increases, each method decreases more slowly on the RVI-38 dataset compared to the MINI-RGBD dataset, reflecting the stronger robustness brought by training the model on a larger dataset. In addition, we are seeing that the accuracy of our system shows a slower decreasing trend under different noise levels, which represents the stronger stability and robustness of our system.

G. Qualitative Analysis

Fig. 4 visualises the interpretability of proposed attention module by presenting the attention value of each joint among all leave-one-out cross-validations on each dataset. We observe that the attention value of 'Right Knee','Left Knee','Right Wrist' and 'Left Wrist' is significantly higher than other joints on both datasets. It indicates our system pays more attention to the movement frequencies of infants' knees and wrists, which is convincing since the movements of those joints have the most significant frequency change in the video recordings. In addition, the frequency range of 'Right Eye', 'Left Eye', 'Right Ear' and 'Left Ear' is lower than other joints significantly. One possible reason is that the self-occlusion (e.g. infant turns head) brings the noise to Openpose estimation of these joints, so that the attention

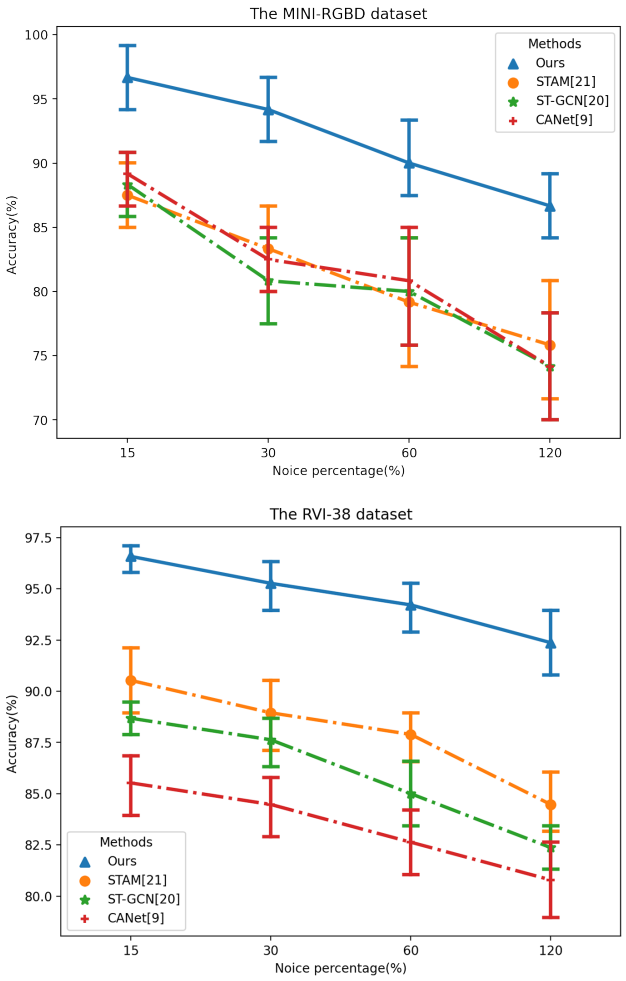

Fig. 3. The robustness test compared with the state-of-the-art DNN methods. The short vertical bar of each method in different noise-level denotes the accuracy range between the first quartile and third quartile among all cross-validations. The line between each bar is linked by the mean accuracy value.

module of our system lowers the weights to filter the noisy data. Besides, we notice that the attention weight range of most of the joints in the RVI-38 dataset is larger than those in the MINI-RGBD dataset. It could be caused by more information from the larger dataset.

V. CONCLUSION

In this work, we propose a novel interpretable frequency attention informed graph convolutional network to predict cerebral palsy infants. We design a binning module for CP data to increase the weight of the low-to-mid frequency data to improve the CP prediction performance, which is adaptable for the videos with a frame rate between 24 FPS to 60 FPS and suitable for both DNN or machine learning-based classification model. Furthermore, we propose a frequency attention module to further improve the prediction performance and visualise the important joints that the network considers in CP prediction. Experimental results show the importance of low-to-mid frequency data and the effectiveness and robustness of our system in in supporting

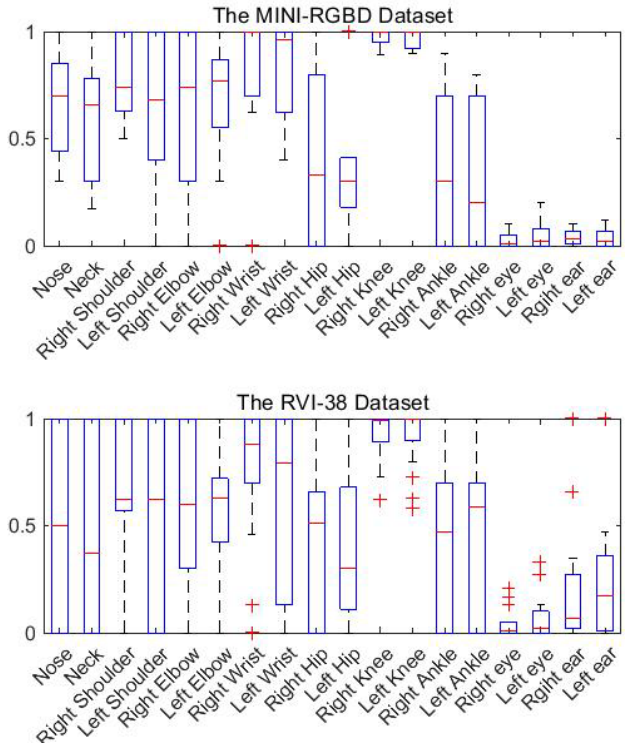

Fig. 4. The visualization of attention weights of different joints among all cross-validations on each dataset.

the prediction of CP non-intrusively, and provides a way for supporting the early diagnosis of CP in the resource-limited regions where the clinical resources are not abundant. Future work is scheduled to apply our methods to a more extensive clinical dataset. In addition, an interesting future work could be using the transformer to interpret the frequency features better.

VI. Compliance with Ethical Standards

The collection of the RVI-38 dataset has been ethically approved by 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). The MINI-RGBD dataset used in this study is made open access by Fraunhofer IOSB [8], which had ethical approval.

References

[1] 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 and developmental disabilities research reviews*, vol. 11, no. 1, p. 61—67, 2005.

[2] D. Graham, S. P. Paget, and N. Wimalasundera., "Current thinking in the health care management of children with cerebral palsy," *Medical Journal of Australia*, vol. 210, no. 3, pp. 129–135, 2019.

[3] S. Orlandi, K. Raghuram, and C. R. S. *et al.*, "Detection of atypical and typical infant movements using computer-based video analysis," in *IEEE EMBC*, 2018, pp. 3598–3601.

[4] E. A. F. Ihlen, R. Støen, and L. Boswell.*et al.*, "Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study," *Journal of Clinical Medicine*, vol. 9, no. 1, 2020.

[5] A. Stahl, C. Schellewald, and O. S. *et al.*, "An optical flow-based method to predict infantile cerebral palsy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 4, pp. 605–614, 2012.

[6] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1225–1234, 2016.

[7] 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. 51 582–51 592, 2020.

[8] N. Hesse, S. Pujades, and J. t. Romero, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018.

[9] 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 *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4.

[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv e-prints*, p. arXiv:1812.08008, Dec. 2018.

[11] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. S. Schroeder, "Computer vision for medical infant motion analysis: State of the art and RGB-D data set," in *ECCV 2018 Workshops*, 2018.

[12] K. D. McCay, P. Hu, H. P. H. Shum, W. L. Woo, C. Marcroft, N. D. Embleton, A. Munteanu, and E. S. L. Ho, "A pose-based feature fusion and classification framework for the early prediction of cerebral palsy in infants," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 1–1, 2021.

[13] 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 *IEEE EMBC*, 2019, pp. 5469–5472.

[14] 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. 42 314–42 324, 2021.

[15] 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. 94 281–94 292, 2021.

[16] F. Ferrari, G. Cioni, and C. t. Einspieler, "Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy," *Archives of Pediatrics and Adolescent Medicine*, vol. 156, no. 5, pp. 460–467, 2002.

[17] C. Einspieler, R. Peharz, and P. B. Marschik, "Fidgety movements – tiny in appearance, but huge in impact," *Jornal de Pediatria*, vol. 92, no. 3, Supplement 1, pp. S64–S70, 2016.

[18] L. Bluestein, "A linear filtering approach to the computation of discrete fourier transform," *IEEE Transactions on Audio and Electroacoustics*, vol. 18, no. 4, pp. 451–455, 1970.

[19] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.

[20] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *AAAI Conference on Artificial Intelligence*, 2018.

[21] 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 Journal of Biomedical and Health Informatics*, pp. 1–1, 2021.

[22] K. D. McCay, E. S. L. Ho, D. Sakkos, W. L. Woo, C. Marcroft, P. Dulson, and N. D. Embleton, "Towards explainable abnormal infant movements identification: A body-part based prediction and visualisation framework," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4.

[23] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[24] 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.