Human Movement Disorders Analysis with Graph Neural Networks

Haozheng Zhang

This dissertation is submitted for the degree of Doctor of Philosophy

Supervisors: Dr Hubert P.H. Shum & Dr Yang Long

Department of Computer Science

The University of Durham

United Kingdom

October 2023

Abstract

Human movement disorders encompass a group of neurological conditions that cause abnormal movements. These disorders, even when subtle, may be symptomatic of a broad spectrum of medical issues, from neurological to musculoskeletal. Clinicians and researchers still encounter challenges in understanding the underlying pathologies. In light of this, medical professionals and associated researchers are increasingly looking towards the fast-evolving domain of computer vision in pursuit of precise and dependable automated diagnostic tools to support clinical diagnosis. To this end, this thesis explores the feasibility of the interpretable and accurate human movement disorders analysis system using graph neural networks.

Cerebral Palsy (CP) and Parkinson's Disease (PD) are two common neurological diseases associated with movement disorders that seriously affect patients' quality of life. Specifically, CP is estimated to affect 2 in 1000 babies born in the UK each year, while PD affects an estimated 10 million people globally. Considering their clinical significance and properties, we develop and examine the state-of-the-art attention-informed Graph Neural Networks (GNN) for robust and interpretable CP prediction and PD diagnosis.

We highlight the significant differences between the human body movement frequency of CP infants and healthy groups, and propose frequency attention-informed convolutional networks (GCNs) and spatial frequency attention based GCNs to predict CP with strong interpretability. To support the early diagnosis of PD, we propose novel video-based deep learning system, SPA-PTA, with a spatial pyramidal attention design based on clinical observations and mathematical theories. Our systems provide undiagnosed PD patients with low-cost, non-intrusive PT classification and tremor severity rating results as a PD warning sign with interpretable attention visualizations.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my principal supervisor Dr. Hubert P. H. Shum, for supporting me in every way possible throughout my studies and life. He has provided me with constant encouragement and led me through the field of human motion analysis and deep learning. I thank him for his incredible insight, his invaluable advice, his endless support, and perhaps most importantly his enduring patience.

I would also like to express my deepest gratitude to my parents who have provided every possible material and spiritual support for my study. Another significant person who I would like to thank to is my partner. My achievement cannot be apart from their continued love and encouragement.

I would also like to thank Dr. Edmond Ho for his continuous support, guidance and helpful advice. Furthermore, I appreciate that I have chance to collaborate the projects with Dr Silvia De Din and my colleagues Xiatian Zhang & Ziyi Chang. They are really professional, creative, and helpful.

Finally, thanks also to my fellow PhD students from the Durham University, for their valuable experience sharing and encouragement during my studies and the many helpful and constructive discussions we have had. In particular, I would like to thank Shuang Chen, Ruochen Li, Li Li, Qiaotan Qiu, Manli Zhu, Xiaotang Zhang, Mridula Vijendran and Ruizhi Liu for their tremendous technical support and idea sharing.

List of Publications

- **Haozheng Zhang**, Edmond S. L. Ho and Hubert P. H. Shum, "CP-AGCN: Pytorch-Based Attention Informed Graph Convolutional Network for Identifying Infants at Risk of Cerebral Palsy," Software Impacts, vol. 14, pp. 100419, Elsevier, 2022.

- **Haozheng Zhang**, Edmond S. L. Ho, Xiatian Zhang and Hubert P. H. Shum, "Pose-Based Tremor Classification for Parkinson's Disease Diagnosis from Video," in MICCAI '22: Proceedings of the 2022 International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 489-499, Singapore, Singapore, Springer, 9 2022.

- **Haozheng Zhang**, Hubert P. H. Shum and Edmond S. L. Ho, "Cerebral Palsy Prediction with Frequency Attention Informed Graph Convolutional Networks," in EMBC '22: Proceedings of the 2022 International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1619-1625, Glasgow, UK, IEEE, 7 2022.

- **Haozheng Zhang**, Edmond S. L. Ho, Xiatian Zhang, Silvia Del Din and Hubert P. H. Shum, "Pose-based tremor type and level analysis for Parkinson's disease from video," International Journal of Computer Assisted Radiology and Surgery, https://doi.org/10.1007/s11548-023-03052-4, 2024.

- Edmund J. C. Findlay, **Haozheng Zhang**, Ziyi Chang and Hubert P. H. Shum, "Denoising Diffusion Probabilistic Models for Styled Walking Synthesis," in MIG '22: Proceedings of the 2022 ACM SIGGRAPH Conference on Motion, Interaction and Games, Guanajuato, Mexico, ACM, 2022.

- Xiatian Zhang, Sisi Zheng, Hubert P. H. Shum, **Haozheng Zhang**, Nan Song, Mingkang Song and Hongxiao Jia, "Correlation-Distance Graph Learn-

ing for Treatment Response Prediction from rs-fMRI," in ICONIP '23: Proceedings of the 2023 International Conference on Neural Information Processing, Changsha, China, Springer, 11 2023.

- Ziyi Chang, Edmund J. C. Findlay, **Haozheng Zhang** and Hubert P. H. Shum, "Unifying Human Motion Synthesis and Style Transfer with Denoising Diffusion Probabilistic Models," in GRAPP '23: Proceedings of the 2023 International Conference on Computer Graphics Theory and Applications, pp. 64-74, Lisbon, Portugal, SciTePress, 2 2023.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human movement disorders refer to a group of neurological conditions that cause abnormal movements. Such abnormal movements are categorized into increased movements (e.g., tremors, dystonia) or decreased movements (e.g., Parkinsonism). Since even subtle movement disorders can be a sign of a variety of conditions ranging from neurological disorders to musculoskeletal problems, accurate diagnosis of movement disorders and potential related neurological conditions is highly relied on experienced neurologists. However, with an aging population and a projected increase in neurological diseases countries all over the world do not have enough clinicians to meet patient needs [3]. For example, there was only one consultant neurologist per 91,175 of the population in the UK based on a neurology workforce survey in 2019 [4]. Hence, medical professionals and associated researchers are increasingly looking towards the fast-evolving domain of computer vision in pursuit of precise and dependable automated diagnostic tools to support clinical diagnosis [5].

Automated human action recognition (HAR) has been a rapidly evolving research field for several years [6]. By identifying and classifying human actions, it can be applied in surveillance systems to detect suspicious or dangerous activities, autonomous driving systems to predict pedestrian behavior, and other applications like human-computer interaction, video retrieval, and sports analysis. In addition,

HAR provides healthcare monitoring solutions to ensure the safety of patients or the elderly through risky activity detection or prediction, such as fall prediction and other irregular human movements detection. Specifically, human pose estimation algorithms provide informative 2D or 3D skeletal data from RGB videos or depth maps. These representations are less affected by backgrounds, relatively robust to occlusions, easy for visualization and interpretation, and cost-effective both in data acquisition and computation, such that they are suitable for real-world clinical applications.

For the skeletal data mentioned above, traditional deep learning methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), encounter difficulties in processing non-Euclidean structured data (e.g., graphs or manifolds) due to the lack of the inherent flexibility to handle the variability of node-wise connections and distances. This prompted researchers to propose graph neural networks (GNNs) specifically designed to learn graph structure data. Specifically, GNN is designed to process graph data composed of nodes and edges, and capture the topology in the data and the complex relationships between nodes. Since the concept of GNN was proposed, GNNs have achieved remarkable results in various fields, such as social network analysis [7], biomedical engineering [8], transportation networks [9], and recommendation systems [10], providing a new perspective for processing and understanding graph-structured data.

In this thesis, we aim to develop state-of-the-art GNNs for analyzing the graph-structure human pose features extracted from video recordings to develop automated disease diagnosis systems. We further complement our system with additional clinical guidance to make them more accurate, robust, and whose decisions can be explained by humans. We wish such automated diagnostic tools could not only be used to assist clinicians in making more comprehensive and precise diagnoses but also offer low-cost diagnostic support for regions with limited clinical resources.

1.1 Research Challenges and Methodology Overview

This section introduces two critical human movement-based diseases we analyze in this thesis: cerebral palsy (CP) and Parkinson's disease (PD) analysis. Although CP and PD are distinct neurological disorders, the movement patterns of their patients show significant differences from those of healthy people, making them suitable for classification using neural networks. Our research found that the framework of extracting pose features from patient video records and then using a GNN-based classification model to diagnose diseases related to human movement demonstrated reliable interpretability and accuracy. In the following subsection, we first discuss the motivations and challenges associated with each disease and then present an overview of the methodologies used for their analysis.

1.1.1 Cerebral Palsy Prediction

Robust and interpretable identification of infants at high risk of cerebral palsy (CP) is critical for early intervention. CP is one of the most prevalent physical disabilities affecting children and occurs in around 2 out of 1,000 live births [11]. In general, the average diagnosis period of CP is formulated to families when the child is approximately 11 months old, but the diagnosis may be delayed until 24 months for those with milder symptoms [12]. As intervention programs are required to commence before the age of 6 months [13], this may result in late intervention and irreversible harm to the patient's life [14].

To detect whether an infant is at high risk for CP, clinical methods prefer a combination of standardized tools in conjunction with clinical history. One of the most predictive tools is the video-based General Movements qualitative Assessment (GMA) [15]. General movements (GMs) are spontaneous movements and involve all body parts. They emerge during early fetal life and disappear when goal-directed motor behavior emerges around 4–5 months corrected age (CA). The form of typical

GMs changes as a result of developmental transformations of the nervous system. In the last phase, at 2 to 5 months CA, GMs have a 'fidgety' character. Fidgety GMs (FMs) occur irregularly all over the body and consist of a continuous stream of tiny elegant movements. During each phase, typical GMs are primarily characterized by complexity and variation [16]. GMs are considered abnormal when their complexity is reduced. However, GMA requires a high level of expertise in the assessors, which impedes the application in broad clinical practice [17]. In pediatric practice, there's a growing need for an automated, user-friendly screening tool as a substitute for expert-dependent GM video ratings [18]. Therefore, for broad applicability, this computer-aided diagnostic tool should be sensitive, accurate, affordable, and permit free movements [19].

Although existing pose-based systems achieved great progress in CP prediction [20, 21, 22], most of the existing models' performance was constrained because they predominantly focus on spatial pose features, which cannot fully represent complex human movements. In addition, interpretability is a critical factor for the clinical validation and practical application of these systems [23]. Despite its importance, there has been limited research on developing methodologies to visualize the detailed decision-making process within complex deep learning frameworks [20]. Effective visualization techniques, such as attention maps or feature attribution methods, are vital for medical applications, as they can provide insights into the model's reasoning, thereby increasing clinician confidence and facilitating the adoption of computer-aided tools in real-world healthcare settings.

To address the above key challenges, we design our CP prediction systems based on two key observations: (i) clinicians indicates a significant difference between CP infants' frequency of human movement and that of the healthy group, specifically, Rahmati et al. [24] found that compared 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; (ii) we found that the infants' joint position data in the high-frequency domain is mainly caused by data noise, such as the mis-

detected joint position by the pose extraction algorithm. Furthermore, the theory of time-frequency consistency [25] demonstrates that frequency domain information is complementary to the spatial domain. Based on these, we hypothesize that modeling CP in the frequency domain could lead to more informative representation learning compared to the spatial domain.

In our first CP prediction work [26] (Chapter 4), we propose a frequency-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 [1] to extract the human joint position data from the RGB 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 less-informative high-frequency movements in the whole video sequence for improving CP prediction accuracy. We propose to investigate the use of GNN in the analysis of human movement diseases, such as CP, since the graph-based structure with nodes and edges offers enhanced capacity and interpretability for modeling human pose data. In addition, the existing publicly available dataset for CP prediction, MINI-RGBD [27], only consists of a limited number (i.e., 12) of synthetic infant body movement video recordings, resulting in network training and system evaluation that lack strong robustness. Therefore, our team has collected the RVI-38 dataset as part of routine clinical care at the Royal Victoria Infirmary (RVI) in Newcastle upon Tyne, UK. It includes 38 RGB video sequences of different infants aged between 12 to 21 weeks, with an average video length of 3 minutes and 36 seconds. We make the pose data of the RVI-38 dataset available to the community to encourage research in this field.

During the above CP prediction work, we further identified three limitations in CP prediction systems. First, existing studies lack a holistic framework to learn both spatial and frequency information [26, 28], where spatial information could be complementary to frequency information, and both contribute to a more comprehensive understanding of CP. Because analysis in the frequency domain primarily

focuses on identifying specific patterns, such as abnormal muscle contractions and rhythmic movement, while spatial domain analysis provides insights into the general physical movements and postures of individuals. Second, while interpretability is crucial for clinical validation and real-world application [23], existing work has limited focus on visualizing the decision process of complex deep learning [20, 26]. Third, whilst the efficacy of pose-based methods is intrinsically linked to the quality of pose features, the de facto pose extraction method, OpenPose, results in less accurate body joints, particularly during self-occlusion [29, 26].

To this end, we further propose a novel two-stream Spatial and Frequency Attention-based Graph Convolutional Network (SAFA-GCN) to fuse the spatial pose and the movement frequency features for robust CP prediction (Chapter 5). Our SAFA-GCN includes a spatial attention module, a masked-frequency attention module, and a clipping-and-fusion method to improve model prediction accuracy and interpretability, facilitating the visualizations of significant human joints, frequency bands, and time ranges during CP prediction. Finally, we supplement CP datasets with new and more accurately extracted pose features.

1.1.2 Parkinson's Disease Analysis

We transfer and validate our research insights and methodologies from the CP prediction project into Parkinson's disease (PD) analysis, as PD is the second most common progressive neurological disease affecting a wider population, with an estimated 10 million people worldwide [30]. It is characterized by the loss of dopaminergic neurons within the substantia nigra region of the brain, resulting in motor dysfunction [31]. Existing PD diagnosis is mainly based on the clinical assessment of PD symptoms, medical history, l-dopa and dopamine responses [32]. The clinical diagnostic accuracy is approximately 73%-84% [33], and may be affected by medical experts' subjective opinions and experiences. An automatic, efficient, and interpretable PD assessment system would support clinicians in making more robust diagnostic decisions.

Recent research in PD diagnosis with machine learning using human-centric visual, audio, and movement features has shown promising results. Models based on neuroimaging [34] and cerebrospinal fluid biomarkers [35] provide an accurate diagnosis but are costly and intrusive, making them unsuitable for large-scale pre-diagnosis. Non-intrusive methods with speech [36] are limited by their generalizability due to the significant difference in language and pronunciation for patients from different geographical areas. Although gait disturbance is not typically the primary symptom of early-onset PD [37, 38], over 70% of these patients exhibit at least one form of tremor [38]. Hence, identifying Parkinson's Tremor (PT) is seen as a more generalizable approach for assisting in early PD diagnosis. To date, hand tremors-based studies mostly rely on wearable sensor data [39]. However, the use and set-up of wearable technology may be time and resource-consuming [39]. Video-based analysis with consumer-grade cameras is preferable as a more cost-effective solution without disrupting the natural behavior of the participants.

To this end, we propose to support PD diagnosis by classifying PT since it is one of the most predominant symptoms of PD with strong generalisability. Different from other computer-aided time and resource-consuming PT classification systems that rely on wearable sensors, we propose the first GNN-based PT analysis system, SPA-PTA, to provide undiagnosed patients with low-cost PT classification and tremor severity estimation results as a PD warning sign via only consumer-grade non-intrusive video recordings. In addition, we propose a novel attention module with a lightweight pyramidal channel-squeezing-fusion architecture to extract relevant PT information and filter the noise efficiently. This design aids in improving both classification performance and system interpretability. Our solution outperforms existing ones in PT analysis, achieving 91.3% accuracy and 80.0% F1-score in PT classification, 76.4% accuracy and 76.7% F1-score in tremor rating classification.

1.2 Summary of Contributions

Our research demonstrates the effectiveness of incorporating clinical knowledge into existing deep neural network (DNNs), specifically, GNN frameworks for improving system accuracy and robustness. In addition, we flexibly apply and develop attention mechanisms to interpret CP prediction and PT analysis. Specifically, our contributions can be summarized as follows:

- In the CP prediction project, we propose two attention-informed GNNs to accurately predict CP by only using the pose features extracted from non-intrusive consumer-grade RGB videos.
- We develop the first frequency attention informed CP prediction GNN, namely FAIGCN, by modeling frequency features to support CP prediction and interpret it in the frequency domain. In addition, we designed a frequency-binning module that can be applied to machine learning networks for videos with different frame rates to improve the CP prediction performance.
- We further improve the CP prediction performance by proposing a novel two-stream GNN, namely SAFA-GCN, to fuse complementary spatial poses and movement frequency features, validated with MINI-RGBD and RVI-38 using a more robust evaluation protocol.
- Our final CP prediction solution, SAFA-GCN, further improves both the interpretability and prediction accuracy by following designs:
  (1) Spatial-wise and frequency-wise attention modules allow the visualizations of which joints and frequency bands contribute to a prediction.
  (2) A clipping-and-fusion method that analyzes individual temporal clips and fuses the results for a final prediction, allowing the visualization of when the movements are significant for making a prediction.

- We supplement the MINI-RGBD and RVI-38 datasets with new and more accurately extracted posture features, and conduct a comprehensive benchmark analysis on leading methods, demonstrating a consistent performance enhancement. We make the pose data of these datasets available to the community to encourage the research in this field.
- In the PD analysis project, we propose a GNN, namely SPA-PTA, to diagnose PD through PT classification and provide tremor severity estimation.
- We propose a novel attention module with a lightweight pyramidal channel-squeezing-fusion architecture to capture the self, short and long-range joint information specific to PT and filter noise. This design aids in improving both classification performance and system interpretability.
- We evaluate the leading systems via a more challenging individual-based leave-one-out cross-validation to improve system robustness, our SPA-PTA outperforms existing ones in PT analysis, achieving 91.3% accuracy and 80.0% F1-score in PT classification, 76.4% accuracy and 76.7% F1-score in tremor rating classification.
- Our works demonstrate the effectiveness and efficiency of computer-assisted technologies in supporting the CP prediction and the diagnosis of PD non-intrusively. Our systems provide the CP and PT classification warning sign for supporting the CP prediction and the diagnosis of PD in resource-limited regions where the clinical resources are not abundant.
- We open our source code of each project for validation and encourage further development in related areas.

1.3 Thesis Structure

The structure of the remainder of the thesis is as follows: In Chapter 2, we first provide a comprehensive literature review on computer-aided CP prediction and

PD analysis methods. Then, we discuss two critical deep learning techniques that highly effective and relevant to CP prediction and PD analysis tasks: the attention mechanism and GNNs.

In Chapter 3, we cover the technical details of two major deep learning techniques in this thesis: the attention mechanism and Graph Convolutional Networks.

In Chapter 4, we provide the initial insight and research outcome about developing an interpretable GNN with an attention mechanism for CP prediction. We also highlight the significant differences in joint motion frequencies between CP infants and healthy groups, which can be effectively learned by frequency-informed GCNs. In addition, we further identified three limitations in existing CP prediction systems, which will be addressed in Chapter 4.

In Chapter 5, we provide full details of our improved CP prediction model, SAFA-GCN, with higher accuracy, stronger robustness, and better interpretability. We also introduce a more robust evaluation protocol and corresponding performance benchmarks to aid future research in this area.

In Chapter 6, we provide details of the proposed PT classification and tremor severity estimation framework with novel GNN design to support PD analysis. We provide information related to qualitative visualization results of the experimental performance and clinical significance of the proposed method.

In Chapter 7, we conclude the research outcome of my PhD study and then provide a comprehensive discussion of my research experience, research limitations, and future directions in this research area.

Chapter 2

Literature Review

In this chapter, we introduce some background knowledge of the two target diseases (i.e., CP and PD) in the thesis, along with the deep learning methods that are applied or can be transferred into the analysis of these two diseases. We first review the mainstream clinical infant CP assessment tool - the general movement assessment. Then, several existing computer-aided CP automatic prediction methods are introduced. Besides, we provide an overview of various machine learning-based methods for PD detection. After that, we specifically discuss two key deep learning techniques that highly effective and relevant to CP prediction and PD analysis tasks: the attention mechanism and GNNs.

2.1 Cerebral Palsy Prediction

This section introduces the clinical aspects and computer-aided perspectives of CP. It covers its prevalence, prognosis, associated developmental milestones, and the impact on individuals diagnosed with CP. Additionally, the section discusses the primary diagnostic tool for predicting CP, the General Movement Assessment (GMA), along with existing methods for its automated prediction. Furthermore, we introduce the current frequency analysis techniques for understanding infants' movements and predicting CP, emphasising the research gaps that merit further investigation.

2.1.1 Cerebral Palsy and the General Movement Assessment

Cerebral palsy (CP) is an umbrella term that covers a variety of persistent neurological conditions. Essentially, CP mainly impairs the patient's mobility, muscle tone, posture, and coordination, such that CP is considered a human movement disorder. In addition to these cardinal symptoms, infants with cerebral palsy may experience many other challenges, including dysphagia and speech articulation problems, hearing deficits, vision problems, epilepsy, gastroesophageal reflux disease, and learning disabilities [40]. CP symptoms vary significantly in severity, with some people experiencing relatively mild symptoms such as motor impairments, while others are severely disabled and face significant challenges in undertaking daily activities.

CP stands out as a significant physical disability in children, presenting itself in about 2 per 1,000 live births [11]. Typically, families are alerted to a CP diagnosis around the child's 11th month. However, in cases showcasing less overt symptoms, this diagnostic timeframe could extend to 24 months [12]. Considering that interventions ideally start by the 6th month [13], delays can lead to missed therapeutic windows, potentially introducing irreversible consequences for the patient's future [14]. Therefore, early diagnosis and intervention are clinically considered the paramount part of treating CP.

For a number of years, the pursuit of a reliable early diagnosis for CP has been at the forefront of medical research. Multiple physical examinations have shown potential in identifying early signs of the CP condition by focusing on infants' muscle tone, engagement, and coordination. Additionally, they consider the precision and intricacy of infants' spontaneous movements during specific developmental milestones. Among these physical examinations, the Prechtl's General Movements Assessment (GMA) [15] achieves the most prominent accuracy and reliability. The GMA is a non-invasive physical assessment tool for identifying neurological anomaly conditions that could potentially result in CP. GMs involve the whole body,

mainly in different sequences of neck, arm, trunk, and leg movements [16]. Around 6-9 weeks after full term, a specific type of GMs known as "fidgety movements" (FMs) gradually occurs, replacing the previously observed "writhing movements", and persists until approximately 16-20 weeks [15]. FMs refer to subtle neck, trunk, and limb movements in various directions and at different accelerations [41]. [15] proposed that infants with a compromised nervous system exhibit a deficiency in FMs, which were associated with CP.

Existing studies have supported the argument of the high relationship between FMs and CP. Ricci et al. [42] proposed that lack of FMs can be considered a crucial indicator for identifying abnormal GMs, thus helping to predict CP. In addition, a systematic review incorporating 47 studies concluded that GMA during the fidgety period exhibits the highest sensitivity (97%) and specificity (89%) among the currently available CP prediction methods [43]. Moreover, as most infants were no longer hospitalized during the fidgety period, the authors encouraged parents to utilise technologies such as smartphone applications to facilitate remote capturing of infant movements during this critical age. This perspective supports the development of video-based CP prediction systems.

2.1.2 Automated CP Prediction Systems

While GMA has been extensively proven as an accurate and reliable non-invasive tool for predicting CP, training a proficient GMA assessor still requires substantial investment. Therefore, several studies in the past decade have focused on developing automated CP prediction methods based on GMA. Pioneering research in this field [44] introduced a method that utilized differences between successive frames to generate a depiction of infant movements for abnormal movement classification. However, this method relies on the difference of images as feature representations, making it heavily affected by self-occlusion issues and potential camera motion constraints. After that, Orlandi et al. [45] used the Large Shift Optical Flow (LDOF) to track the baby's movement and gain speed. Before extracting

features for classification, they calculated the displacement of each pixel within ten frames. In addition, they binary classified the extracted features to determine normal or abnormal general movements using multiple classifiers. [46] proposed a method for early prediction of CP based on GMA theory with RGB videos. They explored human pose recognition in a supine position based on RGB-D videos and applied it to auto-GMA. Specifically, they employed the pose estimation method on RGB images to achieve the infant full-body 2D key points. By combining the depth information, the 3D movement of the infant in a supine position is obtained. Then, they achieved the infant's movement complexity index by extracting the infant's whole-body movement characteristic. However, the robustness of the extracted features in these studies still significantly suffers from noise introduced by occlusion, drift, and susceptibility to unrelated movements. For example, the optical flow-based feature extraction method employed in [45] faces challenges in addressing the incomplete data issue caused by occluded body parts, hindering accurate motion reconstruction.

Conventional machine learning-based CP prediction methods were mainly in 2D space. Das et al. [47] proposed a machine learning-based algorithm by using KAZE points to track infant kicking and collect kinematic data. Each type of movement was classified by computing unique feature criteria and learning motion models using support vector machines (SVM). Ihlen et al. [48] presented a machine learning model, namely the Computer-based Infant Movement Assessment (CIMA) model, for clinically feasible early CP prediction based on infant video recordings. The CIMA model was designed to use time-frequency decomposition of the movement trajectories of the infant's body parts to assess the proportion of movements related to the risk of CP. A linear discriminant analysis (LDA) was used to classify common movements in children with and without CP. However, infant movements have complex temporal patterns [46], which requires the use of more advanced computing techniques and more predictive feature learning, such as deep learning models with technical designs for representation learning that can

handle temporal dynamics and nonlinearity.

In the past few years, deep learning frameworks with pose estimation processes have been applied to automated CP prediction systems, yielding promising results. McCay *et al.* [29] explored the feasibility of extracting pose-based features from video sequences to automatically classify infant body movement into two categories: normal and abnormal. They further 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 approximately 92%. On the same dataset, [49] achieved similar CP prediction performance by incorporating a SENet [50] based CNN, significantly enhancing the interpretability. Additionally, [51] utilized 3D graph convolution to extract spatial-temporal information, followed by an online detection process and an unsupervised pseudo-label generation process to augment the data and improve both model capacity and performance. However, the robustness and generality of their proposed method have not been fully evaluated since the results are obtained from a single small dataset. In addition, the robustness of these methods is constrained as they have only been evaluated on a single dataset with a single data-splitting strategy.

2.1.3 Frequency Analysis in CP Prediction

One of the earliest frequency-based studies of analysing CP is from Rahmati et al. [24]. Specifically, Rahmati et al. [52] first conducted motion segmentation on accelerometer sensor data and RGB video data, and extracted the following three features for classifying CP via the support vector machine(SVM) classifier: areas beyond the standard deviation of the moving average; periodicity; and correlation between trajectories. Based on this, in [24], they performed a frequency-based analysis of accelerometer data, where the movement frequency components were extracted by employing the Fourier transform on the accelerometer sequences. They

provided a result that compared with high-frequency ranges, the middle-to-low frequency range showed more differences between the healthy group and the CP group.

Mills et al. [53] designed a controlled experiment for 11 youth with CP and 16 typically developing (TD) youth aged 7–17 years. Their research indicated that youth with CP behaved like age-matched TD controls at lower frequencies (0.1 and 0.25Hz) during the experiment. In addition, they noted that youth with CP were less able to maintain balance at high oscillation frequencies.

In the study by Stahl et al. [54], a method based on optical flow was introduced, which predicts CP through statistical analysis and pattern recognition of an infant's spontaneous movements. Wavelet frequency analysis was employed to evaluate the temporally correlated trajectory signals found in the optical flow data. However, there were challenges in tracking more significant movements using this optical flow technique, suggesting future analyzes might benefit from videos captured at higher frame rates.

Mccay et al. [28] proposed two histogram-based hand-crafted frequency features, namely Fast Fourier Transform of Joint Displacement (FFT-JD) and Fast Fourier Transform of Joint Orientation (FFT-JO) from the extracted pose sequence from RGB videos. These features provided information on the magnitude of every frequency component extracted from the movement, alongside information on the stiffness, directional shifts, and movement extent linked to the infant's posture. Such insights facilitate a thorough modeling of the infant movement's variability. By classifying CP with an ensemble machine learning classifier, they observed that CP samples showed more low-frequency angular movement at the joints, consistent with their expected characteristics of smoother movements.

Although existing pose-based systems achieved great progress in CP prediction [20, 21, 22, 26, 51], there are three critical limitations that still need to be addressed. Firstly, most of the existing models' performance was constrained by the

fact that only spatial pose features are modelled, which cannot fully represent complex human movements [20, 22, 51]. Those integrating frequency domain analysis lack a holistic framework to learn both spatial and frequency information [26, 28]. Secondly, while interpretability is crucial for clinical validation and real-world application [23], existing work has limited focus on visualising the decision process of complex deep learning [20, 26]. Finally, while the efficacy of pose-based methods is intrinsically linked to the quality of pose features, the de facto pose extraction method, OpenPose, results in less accurate body joints, particularly during self-occlusion [29, 26].

Despite the significant advancements in CP prediction by existing pose-based systems [20, 21, 22, 26, 51], three pivotal limitations persist. First, the performance of existing models is hampered as they primarily model only spatial pose features, inadequately capturing the intricacies of human movement [20, 22, 51]. Approaches that incorporate frequency domain analysis miss an integrated framework to analyze both spatial and frequency details simultaneously [26, 28]. Second, while the system interpretability is paramount for clinical validation and practical application [23], there's a noticeable gap in the existing literature concerning the visualization of decision-making processes within intricate deep learning systems [20, 26]. Lastly, the effectiveness of pose-based techniques heavily relies on the quality of pose features. However, the commonly adopted method for pose extraction, OpenPose, tends to misinterpret body joints, especially during instances of self-occlusion [29, 26].

2.2 Parkinson's Disease Analysis

This section delves into clinical aspects and computer-assisted perspectives of PD. We explore its prevalence, main symptoms, and impact of PD on patients. We also discuss the various major diagnostic methods used for PD detection. In addition, we also introduce different existing computer vision methods for analyzing

PD, including but not limited to Parkinson's gait (PG) classification and Parkinson's tremor (PT) quantification. Lastly, we highlight the research gap in the field of PD diagnosis for further Research.

2.2.1 Parkinson's Disease Background

Parkinson's disease (PD) is a recognisable clinical syndrome with a range of causes and clinical presentations. PD represents a fast-growing neurodegenerative condition; the rising prevalence worldwide resembles the many characteristics typically observed during a pandemic, except for an infectious cause. In most populations, 3–5% of PD is explained by genetic causes linked to known PD genes, thus representing monogenic PD, whereas 90 genetic risk variants collectively explain 16–36% of the heritable risk of non-monogenic PD [55]. Additional causal associations include having a relative with PD or tremor, constipation, and being a non-smoker, each at least doubling the risk of PD.

The diagnosis is clinically based; ancillary testing is reserved for people with an atypical presentation. Current criteria define PD as the presence of bradykinesia combined with either rest tremor, rigidity, or both. In addition, PD symptom is visible during the gait and general posture of patients [56]. Prognostic counselling is guided by awareness of disease subtypes. Clinically manifest PD is preceded by a potentially long prodromal period. Presently, the establishment of prodromal symptoms has no clinical implications other than symptom suppression, although recognition of prodromal parkinsonism will probably have consequences when disease-modifying treatments become available. Treatment goals vary from person to person, emphasising the need for personalised management. There is no reason to postpone symptomatic treatment in people developing a disability due to PD. Levodopa is the most common medication used as first-line therapy. Optimal management should start at diagnosis and requires a multidisciplinary team approach, including a growing repertoire of non-pharmacological interventions. At present, no therapy can slow down or arrest the progression of PD. Therefore, an ef-

ficient and interpretable automatic PD diagnosis system is valuable for supporting clinicians with more robust diagnostic decision-making.

2.2.2 Gait-based and Tremor-based PD Detection

Due to the specificity and significance of Parkinson's gait, some existing methods classify Parkinson's gait to detect Parkinson's disease. Aversano et al. [57] proposed a deep learning method for early detection of PD. This method detects PD and the severity of PD by analysing data extracted from wireless sensors. The identification feature they used was the dynamics of the Vertical Ground Reaction Force (VGRF). They employed two deep neural networks to classify PD subjects from healthy subjects, and the best validation accuracy outperformed the other studies on the same dataset. Alharthi et al. [58] proposed a deep CNN to analyze the data of gait-induced ground reaction force for PD patients and healthy subjects. They further employed layer-wise relevance propagation to make the model's output interpretable and investigate the most significant feature in the spatial-temporal gait ground reaction force signals for PD prediction. Alle and Priyakumar [59] use Linear Prediction Residuals to extract discriminating patterns from gait recordings and then use a 1D convolution neural network with depthwise separable convolutions to perform diagnosis. The proposed network achieved an AUC of 0.91 with a 21 times speedup and about 99% lesser parameters in the model compared to the methods at that time. The ability of the proposed network to identify Parkinsonian gait accurately while being small and fast opens new avenues for it to be deployed in embedded systems with limited memory.

Tremor-based PD detection approaches are mainly using the accelerometer sensor data from different human body parts (e.g. wrist, knee and hip). Kim et al. [60] proposed the first CNN-based framework aiming for estimating clinical Parkinsonian tremor (PT) score by using the data collected from the wrist sensor. They also proposed a 2D image representation for training CNN models by transforming sensor data into the frequency domain using the fast Fourier transform. As

a result, their CNN structure outperformed the previous machine learning-based methods (e.g. random forest, decision tree) by validating their networks on a data set with 143 case sizes. Zhang et al. [61] proposed two CNN-based deep learning algorithms to distinguish PD symptoms with controlled groups by classifying PT class with non-PT class. In addition, they compared CNN on raw sensor data and handcrafted data to show that CNN benefited from training on data decomposed into tremor and activity spectra rather than raw accelerometer sensor data. Oktay and Kocer [62] focused on the differential diagnosis of PT and essential tremor (ET). They employed a convolutional LSTM network to classify the PT and ET, and evaluated their method on a dataset of 23 PD patients and 17 ET patients. The results showed the feasibility of convolutional LSTM on differentiation of tremor. However, they also proposed that their methods showed a 10% lower performance (recall) on postural position than testing position for PT classification.

2.2.3 Automatic Video Analysis on PD

Researchers have invented a few automatic methods to analyze PD videos. The early typical automatic methods primarily focus on the extraction of mainly human skeletal data from patients as the first step, followed by the utilization of machine learning classifiers or statistical methods on these skeleton data to achieve the relevant medical objectives [63, 64, 65, 66, 67, 68]. Recently, deep learning methods are also used for modeling human skeletal data extracted from PD videos to increase model capacity for better performances. Guo et al. proposed a sparse adaptive graph convolutional network to achieve automatic leg agility (LA) assessment for PD patients with an accuracy of 98.85% [69]. Lu et al. used human skeletal data to train a Double-feature Double-motion Network with focal loss [70, 71], which achieved reasonable predictions of MDS-UPDRS scores for PD gait [72, 73]. Besides whole body skeleton information, few studies demonstrated hand skeleton information can also be effectively employed to detect and assess Parkinson's disease (PD) tremors through training with common deep learning

techniques such as Long Short-Term Memory (LSTM) or graph neural network (GNN) [74, 75, 76]. In addition, some studies also extracted the feature from RGB frames and optical flow data directly to train their network and also realise accurate performance for automatic PD severity assessment [77, 78].

Although neuroimagings [79] or cerebrospinal fluid [80] based models perform well, they face a problem of high cost and intrusive. As for the non-intrusive methods, current speech-based models [81] are limited by their generalizability, as the language and pronunciation habits of people in different regions and countries vary significantly. Several studies [82, 83] indicate that gait disturbance is less likely to be the main symptom in patients with early-onset PD, but more than 70% of those patients present at least one type of tremors [84, 85, 83]. Hence, we believe that detecting PD by diagnosing Parkinson's Tremor (PT) is a more generalizable approach compared with other methods. Conventional hand tremors-based studies [86] achieve promising performance by using a deep learning network on wearable sensor data to detect PD. However, using wearable sensors is still time and resource-consuming [86], and requires careful synchronization of data captured from different sensors.

This section discusses the existing graph neural networks (GNNs) and attention mechanisms within the medical domain and explores their potential application in human movement disease diagnosis. Considering the real-world clinical applications, this section is focused on the interpretability and robustness within the attention-informed GNNs.

2.3 Attention Mechanisms

This section introduces several typical attention mechanisms and discusses their potential applications in diagnosing human movement diseases. Given the importance of real-world clinical applications, the focus is on how various attention mechanisms can enhance interpretability of deep learning models and improve

performance.

2.3.1 Spatial Attention Mechanism

. Spatial attention is a mechanism that adaptively selects specific spatial feature regions to focus on [87]. The prevalent method for achieving spatial attention in image data typically involves training a sub-network or constructing a dot product between the query and key, aiming at generating a predictive attention map for the features. [88, 89, 90, 91]. In the case of graph data, Velickovic et al. [92] introduced this approach to the adjacency matrix by proposing the Graph Attention Network (GAT), which assigns varying weights to the relationships between relevant nodes and provides a solution to achieve spatial attention calculation in graph domain.

Recent researches [93, 94, 95] have applied spatial attention mechanisms to the task of human action recognition by modeling skeleton information, demonstrating superior performance in both improving accuracy and enhancing interpretability. In our projects of CP prediction and PD analysis that involves the graph representation learning of the human pose, the attention mechanism in the adjacency matrix can show the importance of different joints (nodes) or bones (edges) in the task of current GNNs. This characteristic is aligns well with the real-world need for interpretable deep learning models in medical research and applications. Despite this, there has yet to be any exploration into the application of spatial mechanism within the analysis of CP and PD video data.

2.3.2 Channel Attention Mechanism.

Channel attention is a channel feature recalibration method that adaptively selects the most informative features while suppressing the less relevant ones. Such a mechanism enables neural networks to identify and focus on the most relevant features. Hu et al. [50] introduces the squeezing-and-excitation network (SENet),

an effective approach for learning channel-wise attention weights, thereby enhancing the generalization capabilities of DNNs. Due to its practicality and innovative approach, the squeeze-and-excitation process pioneered by SENet is often referred to as the foundational form of channel attention [96, 97, 98]. The channel attention mechanism in SENet consists of three processes, squeeze, activation, and scale.

The squeeze process in the channel attention is achieved by applying the global average pooling (GAP) on spatial dimension to capture the global representation features. To capture the channel-wise relationship, the excitation process employs an multilayer perceptron (MLP) with a gated sigmoid activation function. This design learns the non-linear relationship between channels and ensures that the learned relationships are not mutually exclusive. The scale process applies the channel-wise multiplication (i.e., outer product) to obtain the final output.

Existing studies propose various techniques for different tasks in the squeeze step and/or activation step to improve the performance of the model [96, 97, 98, 99]. Based on the channel attention in SENet, SKNet [100] introduced an adaptive kernel selection approach to improve the object recognition performance. Given the significance of joint-wise relationships in analyzing human movement diseases, we have implemented and validated several channel attention designs. These designs aim to enhance model interpretability regarding joint-wise relationships in movements and to improve diagnostic accuracy.

2.3.3 Self-Attention Mechanism.

As one of the most popular and powerful deep learning models in recent years, transformers [101] have achieved great success in the field of natural language processing (NLP) due to their excellent ability to analyze sequential information (especially long-distance information). After that, they significantly promoted the development of computer vision [102]. Therefore, it is essential to explore the application and development of transformers in human movement disease diagnosis

projects.

The core of the initial Transformer architecture is the self-attention mechanism, which computes representations for keys, queries, and values for each element in the input sequence. The value is a matrix where each row represents the hidden state corresponding to an input feature. The key is also a matrix with rows that correspond to the hidden states used to compute attention scores. The query is another matrix that interacts with the key matrix to derive attention scores, indicating the relevance of each element in the sequence to the others. These attention scores are then used to create a weighted combination of the value rows, resulting in a contextually enriched representation for each input element. It implies that the larger the dot product (scaled by the dimensionality of the keys), the larger the attention score will be. A softmax function is applied to normalise the attention weights for each query. The value matrix, is multiplied by these normalized attention scores to create a contextually enriched representation for each input element.

Self-attention mechanisms have predominantly revolutionized Natural Language Processing (NLP), yet their applicability extends to Computer Vision (CV) with notable efficacy. Wang et al. introduced non-local neural networks that conceptualise each pixel location as an individual token, analogous to how words are treated in NLP. These tokens are then utilized to attend to other pixel locations, effectively capturing long-range dependencies within the visual data. This integration of self-attention occurs interstitially between convolutional layers, enhancing the network's capacity to incorporate contextual information across the entirety of the image, which is critical for complex visual understanding tasks.

Over the recent years, various transformer-based methods have presented an exciting performance improvement in different research areas. Ramachandran et al. [102] proposed a full attention model that only employs self-attention layers without any convolutions for image classification. Building upon this framework, Wang et al.[103] introduced a model that adopts positional encodings at the

pixel level and leverages axial attention to sequentially attend to image rows and columns, enhancing the model's ability to capture spatial hierarchies. Furthering this paradigm, Vision Transformers (ViT) [104] had redefined image processing within neural networks by partitioning images into a series of patches. ViT then applied self-attention across these patches layer by layer, offering a more computationally efficient alternative to pixel-level attention calculations. Xing et al. [105] presented a ViT-based model for accurate Alzheimer's Disease diagnosis, incorporating a projection method that projects 3D PET images into 2D fusion images to reduce computational costs.

2.4 Graph Neural Network

This section goes through the research trend of GNNs by introducing several notable GNN methods. Additionally, we discuss the application of existing GNN methods within the expansive field of disease diagnosis.

Whilst conventional deep neural networks (e.g., CNN, RNN and LSTM) have achieved impressive performance in processing structured Euclidean data such as images, videos, and time-series data, they face challenges in handling non-Euclidean data like graphs and manifolds, which require different architectural approaches for effective representation learning [106]. For example, CNNs face difficulties adapting to the irregular and dynamic connectivity inherent in graph-structured data, due to their reliance on fixed adjacency matrices that assume a regular, grid-like topology [107]. Unlike image data, which is represented on a regular grid, graph data lacks this uniform structure. In a graph, each node can have a varying number of neighbors, and the spatial arrangement of these neighbors does not adhere to a fixed pattern. To this end, researches on applying machine learning to graph analysis has been attracted increasing interest because of the significant expressive capability of graphs. For example, graph structures can serve as effective representations of numerous systems in different domains, such as social networks [108], physical

systems [109], knowledge graphs [110], and various other research areas [111].

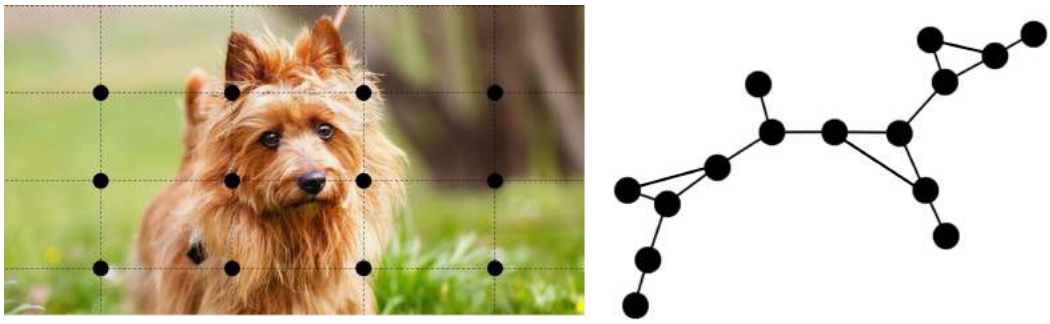


Figure 2.1: An example of image in Euclidean space (Left) and graph in non-Euclidean space (Right).

Graph Neural Networks (GNNs) represent a category of deep learning based methods designed specifically for processing data within a graph domain. Because of their robust performance and solid theoretical foundation, GNNs have recently emerged as a widely applied method for analyzing graph-structured data, especially for classification, link prediction, and clustering tasks. The initial research motivation for the development of GNN came from its use in the 1990s, when Sperduti and Starita [112] used recursive neural networks to analyze directed acyclic graphs. Later, recurrent neural networks and feedforward neural networks were proposed in the studies of [113] and [114] respectively to solve the circular dependency problem. Although these initial attempts were successful, they relied on a process of building a state transition system in a graph and reaching convergence iteratively, a mechanism that limited their scalability and expressive power. The rise of deep learning, especially CNN [115], has promoted the re-exploration of GNNs. Key factors for the success of CNNs include the application of local connections, weight sharing, and multi-layer structures, which are also crucial for solving graph-related problems. However, CNNs are essentially only applicable to regular Euclidean data structures, such as images (2D grids) and text (1D sequences), which can actually be regarded as special cases of graphs. Therefore, extending CNN concepts to graph data should be intuitive. Nonetheless, as shown in Figure 2.1, there are still challenges in defining local convolution filters and pooling operations on

graph data, which limits the extension of CNNs from the Euclidean domain to the non-Euclidean domain.

In recent years, various GNNs have been proposed based on CNN and graph representation learning to model inputs and/or outputs composed of elements and their dependencies. In the following sections, we introduce two primary classifications of GNNs: spectral-based GNNs and spatial-based GNNs. Then, the discussion extends to advanced GNN frameworks that are specifically tailored for applications in medical research.

2.4.1 Spectral-based GNNs

Spectral-based GNNs are founded on the solid spectral graph theory and graph signal processing [116]. They process graph data by first transforming the data into the spectral domain using the graph Fourier transform. Convolution is then applied within the spectral domain, specifically utilizing the eigenvectors of the Laplacian matrix. After convolution, the representations are transformed back to the original domain using the inverse graph Fourier transform.

In the following paragraphs, we introduces several typical spectral-based GNNs with distinct filter designs in the convolution process.

Spectral Network. Bruna et al. [117] proposed a non-spatially localized convolution filter based on a learnable diagonal matrix, but suffered from high computational cost. Henaff et al. [118] further improved this filter to become spatially localized by proposing a parameterization method.

ChebNet. Based on the Chebyshev polynomial based convolution filter approximation theory [119], Defferrard et al. [120] proposed the ChebNet by defining the convolution by multiple localized Chebyshev polynomial. Therefore, this method did not require the computation of the Laplacian eigenvectors.

GCN. Graph Convolution Network (GCN) [121] pioneered the widespread application of convolution operations in the field of graph structure data modeling. This work has primarily addressed two pivotal challenges in convolution-based graph representation learning. Firstly, GCN had countered the issue of overfitting by simplifying the convolution operation to a lower order. Then, they had tackled the problem of vanishing gradients by introducing a novel renormalization technique. It is noteworthy that, due to this renormalization technique, GCNs can also be classified under spatial-based GNNs that would be discussed in the next subsection.

AGCN. Li et al. [122] proposed the Adaptive Graph Convolution Network (AGCN) to learn implicit relationships between different nodes, which had not been modeled effectively in previous GNNs. Specifically, it is achieved by learning the residuals of graph Laplacian. Extensive experiments conducted on various graph-structured datasets had validated the effectiveness of AGCN.

GWNN. Different from the above methods that focus on improving the convolution filter and graph Laplacian, Graph Wavelet Neural Network (GWNN) proposed to replace Fourier transform by the graph wavelet transform. This design is motivated by the benefits of efficient computation of graph wavelet transforms without relying on matrix factorization. Furthermore, due to its sparsity and locality, graph wavelet transform can produce more accurate and interpretable results.

Herein, spectral-based GNNs have a solid theoretical foundation, and several theoretical analyzes have been proposed recently. However, in almost all of the spectral approaches mentioned above, the learned filters are dependent on the graph structure, meaning the filters cannot yet be applied to a graph with a different structure.

2.4.2 Spatial-based GNNs

Different from the spectral-based GNNs that rely on transforming graph data into the spectral domain, spatial-based GNNs leverage various graph topologies to process the convolution operations directly on the graph domain. The main challenge of the spatial-based GNNs is how to effectively define convolution operations with neighborhoods of different sizes and maintain the local invariance in the graph domain.

In the following paragraphs, we introduces several typical spatial-based GNNs with distinct graph topology and convolution designs.

GraphSAGE. GraphSAGE [123] is an inductive framework for generating node embeddings by aggregating node features from their local neighborhoods. This framework has significantly advanced research in spatial-based GNNs, inspiring developments in models such as GCN [121], LCN [124], ST-GCN [125], and GAT [92].

LCN. Ci et al. [124] proposed the Locally Connected Network (LCN) to address the critical limitation of the aforementioned GCN: the fixed weight-sharing mechanism limits the model's capacity for the 3D pose estimation task. In particular, LCN allocates unique convolution filters for distinct nodes, offering a more flexible node representation learning approach.

ST-GCN. Spatial Temporal Graph Convolutional Network (ST-GCN) [125] is a representative graph convolutional neural network for human action recognition task by using human skeleton data. It defines different spatial temporal graph topologies and node partition strategies to achieve effective representation learning via convolutions in both domains.

2S-AGCN. To enhance the utilization of multilevel semantic information for more

informative representation learning, Shi et al. [126] proposed the Two-Stream Adaptive Graph Convolutional Networks (2S-AGCN). This model introduces a two-stream network architecture that integrates bone information (i.e, length and direction) with skeleton information.

GAT. The Graph Attention Network (GAT) [92] pioneered the integration of the attention mechanism within the graph propagation stage. It calculates the hidden states of each node by employing the self-attention mechanism [101] on its neighbors. Additionally, GAT leverages multi-head attention [101] for the calculation of hidden states, subsequently fusing their features to enhance the stability the learning process.

Compared to spectral-based GNNs, spatial-based GNNs offer enhanced flexibility in graph topology design and greater scalability, largely due to their avoidance of expensive spectral-domain transformations or the decomposition of the Laplacian matrix. These advantages have led to the broader application of spatial-based GNNs across various research areas [123, 92, 126, 127]. However, despite these benefits, the development of diverse graph topologies and convolution operations in spatial-based GNNs often encounters challenges related to the lack of solid theoretical foundations. This can make it difficult to understand the underlying reasons for model performance.

2.4.3 GNNs for Disease Diagnosis

In medical diagnostics, the objective is to interpret a patient's symptoms by identifying the underlying diseases [128]. However, according to data from the U.S. healthcare system, patients with serious conditions are particularly prone to misdiagnosis, with an estimated 20% misdiagnosed at the primary care level, and one-third of these errors leading to serious harm [129, 130]. With the fast development of machine learning techniques in divers domain, machine learning-assisted diagnostic technologies are hoped to radically transform the healthcare industry by

leveraging rich patient data to deliver precise and personalized diagnoses [131, 132]. In recent years, GNN-based deep learning systems have primarily been developed and applied for disease diagnosis through medical imaging (MI) and electronic medical records (EMR) data, rather than in analyzing human movement disorder videos [133]. The reason for this preference may be that the former types of data are generally easier to collect during the routine clinical care, have a larger data size, and contain less noise than the latter. Therefore, in the following paragraphs, we mainly introduce GNNs for MI-based disease diagnosis and EMR-based disease diagnosis.

MI-based Diseases Diagnosis. Due to the theoretical support from the theory of brain connectivity [134], GNNs are considered a better alternative to traditional CNNs for brain map analysis. Specifically, GNNs can be designed not only to learn the degree of activation of different brain areas (nodes) but also to understand the connectivity (edges) between them [135, 136].

Parisot et al. [136] were the first to apply graph convolutional networks to the analysis of large-scale brain data, modeling both medical imaging and non-imaging data at the same time. They achieved state-of-the-art results at the time in predicting autism spectrum disorders and Alzheimer's disease across two major datasets. Utilizing quantitative susceptibility mapping (QSM) data, Tang et al. introduced a causality-informed GCN designed to facilitate robust PD diagnosis via the application of the invariant prediction principle and causal interventions [137]. Notably, the features identified by this method are consistent with previous medical research on postural abnormalities associated with PD, indicating that the proposed method is reliable. GNNs also have shown promising performance in diagnosing brain tumors. For example, Song et al. proposed an interpretable structure-constrained GNN (ISGNN) for diagnosing glioblastoma multiforme, showcasing superior interpretability in diagnostic outcomes. By utilizing a metric-based meta-learning approach, they aggregated class-specific graph nodes and concentrated on meta-

tasks related to multiple small graphs, thereby enhancing classification efficacy on small-scale datasets.

EMR-based Diseases Diagnosis. With the increasing prevalence of EMR, disease prediction has gained significant research attention. Various GNNs have been developed to train accurate classifiers that map input predictive signals (e.g., symptoms, patient demographics) to the estimated diseases, aiming to forecast the onset of diseases or diagnosis the diseases.

Sun et al. [138] proposed a GNN-based model for diverse disease prediction that enhances limited EMR data with information from external knowledge bases. Specifically, the proposed model aggregates neighbor node information to generate embeddings that combine both of medical concept and patient record data sources, allowing for accurate disease predictions. Lu and Uddin [139] proposed a GNN-based framework to predict chronic disease by classifying features from latent relationship between patients. Experimental results show that the model achieved approximately 90% accuracy in predicting cardiovascular disease and chronic lung disease. Moreover, Golmaei and Luo [140] proposed DeepNote-GNN to integrate clinical note information and patient network topology, achieving significant improvements in 30-day readmission prediction.

In the previous parts, we discussed the methodologies and characteristics of different attention mechanisms, as well as the researches on GNNs in MI and EMR data modeling for disease prediction, demonstrating the significant potential of these deep learning techniques in the field of disease diagnosis. Although these advances are encouraging, the researches to human movement disorders has been relatively underexplored. Therefore, our study aims to bridge this research gap by focusing on the development of attention-enhanced GNN models specifically tailored for CP prediction and PD analysis.

Chapter 3

Preliminaries

In this preliminary chapter, we cover the technical details of aforementioned two major deep learning techniques in this thesis: the attention mechanism and Graph Convolutional Networks (GCN).

3.1 Attention Mechanisms

Considering the feasibility of improving DNNs' interpretability through different attention mechanisms, we introduce three baseline attention methods in this section: the basic attention mechanism, channel attention mechanism, and self-attention mechanism.

3.1.1 The Basic Attention Mechanism

We follow [141] to present the concept of the basic attention mechanism, which is also widely recognized as the MLP-based attention. Suppose $\mathbf{h} = [\mathbf{h_1}, ..., \mathbf{h_c}, ..., \mathbf{h_C}] \in \mathbb{R}^C$ is a feature map, where C is the number of channels. A basic attention on the channel dimension is computed by:

$$u_c = tanh(Wh_c + b), \tag{3.1}$$

$$\alpha_c = softmax(v^C u_c), \tag{3.2}$$

$$s = \sum_{c=1}^{C} \alpha_c \mathbf{h}_c, \tag{3.3}$$

where $W \in \mathbb{R}^{C \times C}$ is the learnable parameter matrix, and $v, b \in \mathbb{R}^C$, then, $\alpha_c \in \mathbb{R}$ denotes attention weight, $s \in \mathbb{R}^C$, $u_c \in \mathbb{R}^C$. This attention mechanism can be flexibly modified to learn attention maps of different dimensions.

3.1.2 Channel Attention Mechanism

In this subsection, we introduce the channel attention mechanism in SENet [50] that consists of three processes, squeeze, activation, and scale. Suppose a feature map $\mathbf{h} = [\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_C}] \in \mathbb{R}^{H \times W \times C}$, the global average pooling based squeeze process calculate a static $\mathbf{z} \in \mathbb{R}^C$ and the c-th element z_c in $\mathbf{z}$ is computed by:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \tag{3.4}$$

To capture the channel-wise relationship, the excitation process employs an multilayer perceptron (MLP) with a gated sigmoid activation function (seen in Eq.3.5). This design learns the non-linear relationship between channels and ensures that the learned relationships are not mutually exclusive.

$$\mathbf{s} = sigmoid(\mathbf{W}_2 ReLU(\mathbf{W}_1 \mathbf{z})), \tag{3.5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, and r is a hyperparameter of reduction ratio.

The scale process applies the channel-wise multiplication to obtain the final output $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_C]$:

$$\tilde{\mathbf{x}}_c = s_c \mathbf{x}_c. \tag{3.6}$$

3.1.3 Self-Attention Mechanism

The core of the self-attention mechanism involves the computation of representations of keys, queries, and values for each element in the input sequence.

Given a query q_i, the attention score a_{ij} of all keys k_j is calculated by the scaled dot-product:

$$a_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}} \tag{3.7}$$

where d_k is the dimensionality of the keys and queries that remains constant, ensuring that the dot products don't get too large under the softmax function.

It implies that the larger the dot product (scaled by the dimensionality of the keys), the larger the attention score a_{ij} will be. A softmax function is applied to normalise the attention weights for each query. The value matrix, V, is multiplied by these normalized attention scores to create a contextually enriched representation for each input element. The overall self-attention process is summarized as follows:

$$\mathit{Attention}(K, Q, V) = \mathit{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{3.8}$$

with $K = \mathbf{W}_k X, Q = \mathbf{W}_q X$ and $V = \mathbf{W}_v X$, where $\mathbf{W}_k$, $\mathbf{W}_q$ and $\mathbf{W}_v$ are learnable weights matrices for keys, queries and values, respectively, and X is the input matrix.

3.2 Graph Convolutional Network

In this subsection, we introduce the methodology of the representative GNN model, GCN [121]. Suppose a set of graph structure data $G = (V, E)$, where $V = \{x_{i,j} \mid 1 \leq i \leq D, 1 \leq j \leq N\}$ is the set of nodes (or vertices) with N nodes and D features in each node, and E is the set of edges. The propagation in a graph convolution layer can be generalized into a non-linear function:

$$H^{l+1} = f(H^l, A), \tag{3.9}$$

where $H^l \in \mathbf{R}$ is the activation matrix at layer l, A is the adjacency matrix and $f(\cdot)$ is a non-linear function.

An intuitive implementation of Eq.(3.9) is:

$$H^{l+1} = \sigma(AH^l W^l), \tag{3.10}$$

where $\sigma(\cdot)$ is a non-linear activation function and W^l is a trainable weight matrix at layer l. This equation employs the multiplication of the adjacency matrix A with the feature matrix H to denote the aggregation of a node's neighbouring features. Such that the stacking of several hidden layers enables the assimilation of information from multiple levels of neighbouring nodes. However, such a design overlooks the self-influence of a node on itself, and the adjacency matrix A is not normalized, resulting in nodes with many neighbour nodes tending to have greater influence.

To solve the aforementioned limitations, GCN [121] proposed the layer-wise propagation rule Eq.(3.11) by involving the symmetric normalized Laplacian matrix L^{sym}:

$$H^{l+1} = \sigma(L^{sym} H^l W^l) \quad \text{with} \quad L^{sym} = D^{\frac{1}{2}} \hat{A} D^{-\frac{1}{2}} = I_n - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{3.11}$$

where $\hat{A} = A + I_n$ is the adjacency matrix with self-connection by adding the identity matrix I_n, and $D_{i,j} = \sum_j \hat{A}_{i,j}$ such that L^{sym} achieves the required normalization process.

3.3 Conclusion

In this chapter, we have introduced the mathematical foundations and technical details underlying three mainstream attention mechanisms, alongside the baseline GNN model, the GCN. This preliminary introduction helps to understand the specific methodologies and contributions of attention-enhanced GCNs in improving DNNs' interpretability, linking the existing attention and GNN methods to our proposed novel methodologies in Chapters 4, 5, and 6, i.e., attention-informed interpretable GNNs for CP prediction and PD analysis.

Chapter 4

Pose-based Cerebral Palsy Prediction

In this chapter, we provide the initial insights and research outcome of developing interpretable GNNs with attention mechanisms for CP prediction. In addition, we highlight the importance of low-frequency components and redundancy of high-frequency information in infant movements to support further research.

4.1 Introduction

General Movement Assessment (GMA) [15] 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 [142]. 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 [45, 48]. 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.* [54]

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.* [24] 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.

Recent deep learning-based systems achieved impressive performance in CP infants movement prediction. McCay et al. [29] 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 [2]. Zhu [20] 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 [1] to extract the human joint position data from the RGB 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 [24] and our observation. Rahmati *et al.* [24] 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

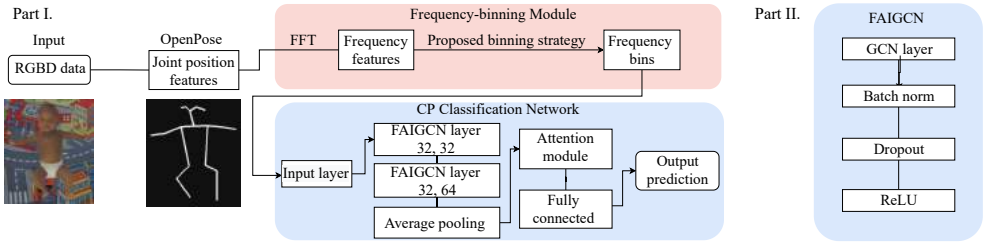


Figure 4.1: The overview of our proposed framework. Part I is the overall network architecture, Part II is the design of each FAIGCN layer.

as the misdetected joint position by Openpose.

In this chapter, we present a phased results of automated CP prediction method. We validate our system on the MINI-RGBD dataset [27] and the RVI-38 dataset [28]. The MINI-RGBD dataset has been widely used for CP classification performance comparison in the previous work [29, 143, 144, 145], 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 videos. Our system achieves state-of-the-art research on two datasets with strong robustness.

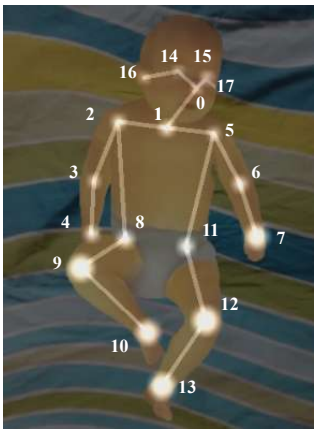


Figure 4.2: An example frame of 18-joints Openpose [1] posture layout for infant kicking in the MINI-RGBD dataset [2]. The size of the light point represents the size of the attention value, that is, the importance of the movement frequency of the joint to our network in CP prediction at that frame.

4.2 Methodology

4.2.1 System Overview

The proposed system consists of two parts (seen in Fig.4.1): (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. 4.2 shows an example of attention visualization.

4.2.2 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 [24], 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 [146]. Previous work [146, 16] has shown that the absence of FMs is an essential distinguishing feature of CP infants from healthy infants. 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.

Fast Fourier Transform (FFT)

We apply Bluestein's FFT algorithm [147], 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:

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-i2\pi kn}{N}}, \quad k = 0, \ldots, N-1, \tag{4.1}$$

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

The Binning Strategy

We propose a feature 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:

$$b_n = \begin{cases} \text{Round}(b_0 \cdot c^n), & \text{if } b_n \cdot c^n < 3; \\ \text{Ceiling}(b_0 \cdot c^n), & \text{if } b_n \cdot c^n \geq 3, \end{cases} \tag{4.2}$$

where b_n is the width of the n^{th} bin, $b_0 = 1$, and c is a controllable hyperparameter designed to maintain the bin width to its original width in low-frequency band, and to increase exponentially when frequency becomes much larger. 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. The threshold of three units is based on our empirical experiments, which are related to video length. 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 4.4, we achieve the best prediction accuracy when $c = 1.00264$ for the 25 FPS videos. The hyperparameter c could be optimized automatically using grid search to achieve the best binning results, as determined by the highest CP prediction performance across 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.

4.2.3 The CP Prediction Network

As shown in Fig. 4.1, we propose a Frequency Attention Informed Graph Convolutional Network (FAIGCN) for CP prediction by classifying low-to-mid fre-

quency band infant movement frequency features with the attention mechanism.

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 [1].

Therefore, we employ a GCN [121] to learn the infant joints dependencies from the pose graph. Inspired by [125], 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, $\{V = v_{b,i} | b = 1, ..., B; i = 1, ..., N\}$ denotes the frequencies of all joints, where $v_{b,i}$ represents the b-th frequency bin of i-th joint. The edge set E includes: (1) the intra-skeleton connection at each frequency band, $\{v_{b,i}v_{b,j} | (i, j) \in K\}$, 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, $\{v_{b,i}v_{(b+1),i}\}$.

The graph convolutional operation of FAIGCN is followed by [121], where the propagation rule between layers can be represented by Eq.(4.3).

$$\mathbf{H}^{l+1} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{A}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{l}\mathbf{W}^{l}\right), \tag{4.3}$$

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.(4.4):

$$\mathbf{v}_k = \sum_{b=1}^{B} \alpha_{b,i} \mathbf{h}_{b,i}, \tag{4.4}$$

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

$$\alpha_{b,i} = \frac{\exp\left(\sigma_n^{'}\left(\mathbf{w}_\alpha^{\mathsf{T}}, \mathbf{z}_{b,i}\right)\right)}{\sum_b \exp\left(\sigma^{'}\left(\mathbf{w}_\alpha^{\mathsf{T}}, \mathbf{z}_{b,i}\right)\right)}, \tag{4.5}$$

$$\mathbf{z}_{b,i} = tanh\left(\mathbf{W}_z \mathbf{h}_{b,i}\right), \tag{4.6}$$

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

$$\sigma_n^{'} = \begin{cases} 1 + \left(\frac{\mathbf{w}_\alpha}{\|\mathbf{w}_\alpha\|}\right)^{\mathsf{T}} \left(\frac{\mathbf{z}_\alpha}{\|\mathbf{z}_\alpha\|}\right) & , \quad n = 1; \\ \mathbf{w}_\alpha^{\mathsf{T}} \mathbf{z}_\alpha & , \quad n = 2, \end{cases} \tag{4.7}$$

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

Implementation Details

As can be seen from Fig.4.1, 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 channel sizes, respectively. This shallow and narrow design is specifically chosen to mitigate the overfitting risk imposed by the limited sizes of the datasets. 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. We adopt cross-entropy loss as the loss function because it is a commonly used and effective loss in binary classification problems. 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.

We release the implementation of our system in PyTorch due to the availability of compatible open-source resources and its good coverage of the required deep learning functionalities. Our code is publicity available for reproduce and further development and the corresponding metadata is shown below:

Nr.	Code metadata description	Please fill in this column
C1	Current code version	v1
C2	Permanent link to code/repository used for this code version	https://github.com/zhz95/CP-AGCN
C3	Permanent link to Reproducible Capsule	https://codeocean.com/capsule/6073072/tree/v1
C4	Legal Code License	MIT License
C5	Code versioning system used	git
C6	Software code languages, tools, and services used	Python
C7	Compilation requirements, operating environments & dependencies	Python 3.7, PyTorth 1.8.10, OpenPose 1.7.0
C8	If available Link to developer documentation/manual	https://codeocean.com/capsule/6073072/tree/v1
C9	Support email for questions	haozheng.zhang@durham.ac.uk

4.3 Dataset and Preprocessing

We verify our models on the Moving INfants In RGB-D synthetic dataset (MINI-RGBD) [27] and RVI-38 dataset.

4.3.1 The MINI-RGBD Dataset

MINI-RGBD was generated by registering and rendering the synthetic Skinned Multi-Infant (SMIL) model [2] 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 [143], 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 [15].

4.3.2 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 video sequences of different infants between 12-21 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.

4.3.3 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 [2], which had ethical approval.

4.3.4 Data Preprocessing

For effective CP predictions, we extract 2D skeleton features from the video sequences. In this chapter, we applied OpenPose [1] for pose estimation since it was one of the most accurate methods for estimating the posture of infants at that time

before the release of the latest version of AlphaPose [148], and it was less sensitive to variations on the appearance [21]. 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 origin and the neck joint with the y-axis and keep the neck joint above the global origin.

4.4 Experiments

Our experiments were run on a PC with Ubuntu 18.04 and an NVIDIA GeForce RTX 3080. Regarding computational efficiency, our low-cost system can achieve a training speed of around 4 frames/second with an NVIDIA GeForce RTX 3080, with only 0.3 frames/second drop compared with our variant without the attention module. It means the total system training time on a 12 video sequences (1000 frames each) dataset is only about 50 minutes, including OpenPose pose estimation. This is considered to be very efficient in deep learning algorithms. During clinical environment inference process, it only needs about 45s for the CP classification of a 33.3s 30 FPS video with an Intel Core i7 CPU (i.e., no GPU needed). It shows that our system is employable for interactive-time prediction in a daily hospital environment with normal computer equipment setting, and is a feasible solution to

support CP early prediction.

4.4.1 Experimental Settings

We conduct the leave-one-out cross-validation among two datasets to evaluate our proposed system. This setting ensures that the prediction system is evaluated on unseen data by utilizing the entire dataset, thereby reducing the risk of overfitting when training with a limited amount of data. The training process strictly follows an early-stop criterion to prevent overfitting. Specifically, if the model's performance on the validation set does not improve for a predefined number of consecutive epochs, the training is halted. Our evaluation metrics are introduced in Sec 4.4.3. We report the best result for each method to be consistent with several related works in literature [54, 16, 24, 145, 28].

4.4.2 Comparing with State-of-the-art Methods

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

- **FCNet** [29]: 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** [29]: 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** [29]: They are two 2D convolutional neural networks, each of them consists of two 2D convolutional layers with differences in the output channel sizes.
- **CANet** [20]: 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) [125]: This is a graph convolutional neural network for human skeleton data (e.g. joint position).
- **STAM** [22]: This is a spatial-temporal graph convolutional neural network with the attention mechanism.
- **Ens-1** [143]: This method uses an ensemble classifier on the fused feature of HOJO2D + HOJD2D (HOJO+D2D) with eight bins.
- **Ens-2** [28]: This method extends [143] by fusing four pose-related features and three velocity-related features.
- **Ens-3** [21]: This method extends [143] by extracting features at limb-level from small video segments to locate abnormal movements spatiotemporally.
- **MCI** [144]: 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.

4.4.3 Evaluation Metrics

We evaluate our system and other state-of-the-art methods by following five metrics in Eq.(4.8): 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) [149] provides a reliable performance metric for imbalanced dataset [150]. The combination of these five metrics comprehensively evaluates the reliability and robustness of the model with consideration for clinical applications [21].

$$\begin{aligned} AC &= \frac{TP + TN}{TP + FN + TN + FP}, \\ SE &= \frac{TP}{TP + FN}, \\ SP &= \frac{TN}{TN + FP}, \\ F1 &= \frac{2TP}{2TP + FP + FN}, \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \end{aligned} \tag{4.8}$$

where True Positive (TP) measures the cases where impaired infants are correctly identified as impaired, while True Negative (TN) indicates unimpaired infants correctly identified as unimpaired. Conversely, False Positive (FP) denotes unimpaired infants mistakenly classified as impaired, and False Negative (FN) refers to impaired infants incorrectly classified as unimpaired. Based on these metrics.

Method	Feature	AC	SE	SP	F1	MCC
FCNet [29]	HOJD2D	91.67	**100.00**	87.50	88.89	83.67
FCNet [29]	HOJO2D	83.33	75.00	87.50	75.00	62.50
Conv1D [29]	HOJO/D2D	83.33	75.00	87.50	75.00	62.50
Conv2D [29]	HOJO/D2D	83.33	75.00	87.50	75.00	62.50
Conv2D [29]	HOJO+D2D	91.67	**100.00**	87.50	88.89	83.67
Conv2D [29]	Pose	83.33	75.00	87.50	75.00	62.50
CANet [20]	Pose	91.67	**100.00**	87.50	88.89	83.67
ST-GCN[125]	Pose	91.67	**100.00**	87.50	88.89	83.67
STAM [22]	Pose	91.67	**100.00**	87.50	88.89	83.67
Ens-1 [143]	HOJO+D2D	91.67	**100.00**	87.50	88.89	83.67
Ens-2 [28]	Velocity	91.67	**100.00**	87.50	88.89	83.67
Ens-2 [28]	Pose*	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Ens-2 [28]	Vel.+Pose*	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Ens-3 [21]	HOJO+D2D	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
MCI [144]	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.

Table 4.1: The comparison with state-of-the-arts on the MINI-RGBD

4.4.4 Comparison with the State-of-the-arts

We report the prediction results on MINI-RGBD and RVI-38 datasets on Table 4.1 and Table 4.2 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 DNNs as CANet outperforms Conv2D-Pose from Table 4.1 and Table 4.2, and CANet outperforms Conv1D/Conv2D, STAM outperforms ST-GCN from Table 4.2.

(3) From Table 4.2, 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 analyze 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 Machine learning-based methods, such as Ens-1, Ens-2, and Ens-3, outperform DNN-based methods with the exception of FAIGCN. Notably, Ens-2 [21], which utilizes an ensemble of classification models (i.e., SVM, LR, LDA, and DT) on the fused hand-crafted features [143, 21], shows performance equal to our FAIGCN on both datasets. On the one hand, this observation shows the superiority of hand-craft features in the classification tasks; On the other hand, it inspires us to explore comparisons of FAIGCN with machine learning-based methods using the same features, seen in the Sec. 4.4.6 below.

Method	Feature	AC	SE	SP	F1	MCC
Conv2D [29]	Pose	81.58	33.33	90.63	36.36	25.85
CANet [20]	Pose	86.84	66.67	90.63	61.54	53.89
ST-GCN [125]	Pose	89.47	66.67	93.75	62.50	60.42
STAM [22]	Pose	92.11	**83.33**	93.75	76.92	72.51
Ens-1 [143]	HOJO+D2D	94.74	**83.33**	96.88	83.33	80.21
Ens-2 [28]	Velocity	94.74	**83.33**	96.88	83.33	80.21
Ens-2 [28]	Pose*	94.74	**83.33**	96.88	83.33	80.21
Ens-2 [28]	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

Table 4.2: The comparison with state-of-the-arts on the RVI-38

4.4.5 Ablation Study

We conduct an ablation study to evaluate whether there is any adverse effect on prediction performance caused by the frequency-binning module (B.) or the attention module (A.). The corresponding results are displayed in Table 4.3. It is noticeable that the frequency-binning module improves the performance of the AC, SP, F1 and MCC metrics in both datasets, showcasing that the frequency-

binning strategy effectively captures and utilizes frequency-related information, which is crucial for enhancing the prediction performance of the model. However, due to the limited size of the MINI-RGND dataset, the contribution of the attention mechanism is not significantly reflected. This result may indicate that the attention module's ability to reweight relevant features does not translate into significant performance gains when the dataset is relatively small, possibly due to insufficient diversity in the training data. Yet, from the larger RVI-38 dataset, it is evident that applying both the attention module and the frequency-binning module leads to significant improvement. These results may suggest that implementing appropriate data augmentation strategies could further enhance the performance improvements achieved by the proposed attention module and frequency-binning module.

4.4.6 Comparison with Machine Learning Methods

To further evaluate the effectiveness of the proposed frequency-binning module and our full model, we employ four machine learning-based classifiers, enhanced by the frequency-binning module, to predict the CP from movement frequency features across 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 [143], Ens-2 [28] and Ens-3 [21] 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 4.4. We observe the effectiveness of using the proposed frequency-binning module as each method outperforms its variant without frequency-binning module, except in the case of SVM in the MINI-RGBD dataset. In addition, we notice that our system outperforms the implemented machine learning-based methods, demonstrating the effectiveness of graph convolutional neural network in dealing with the same features.

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 4.3: The performance of FAIGCN and its simplified variants

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

Table 4.4: The comparison with machine learning based methods and their variant without frequency-binning module

Methods	Bins	AC	SE	SP
K-means cluster binning, K=300	300	0.917	**1.000**	0.875
K-means cluster binning, K=250	250	0.833	0.750	0.875
K-means cluster binning, K=200	200	0.833	0.750	0.875
Equal-width binning, width=3	333	0.917	**1.000**	0.875
Equal-width binning, width=4	250	0.833	0.750	0.875
Equal-width binning, width=5	200	0.833	0.750	0.875
Ours - Frequency binning	299	**1.000**	**1.000**	**1.000**

Table 4.5: Parameter analysis for the binning algorithm

4.4.7 Comparison with Different Binning Algorithms

To evaluate the effectiveness of the insight behind our proposed frequency-binning strategy (i.e., reducing high-frequency noise while preserving low-to-mid frequency information), we compare our method with two commonly-used binning algorithms: K-means clustering binning and equal-width binning, as presented in Table 4.5. The results support our insight and demonstrate the effectiveness of the proposed frequency-binning strategy, since the other two binning algorithms cannot filter out high-frequency noise while retaining low-to-mid frequency features.

4.4.8 Robustness Test

In order to evaluate the robustness of our system and other state-of-the-art DNN-based methods [20, 125, 22], 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. 4.3. The accuracy in the y-axis is the average accuracy among ten leave-one-out cross-validations with ten different training seeds.

From Fig. 4.3, 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 decreas-

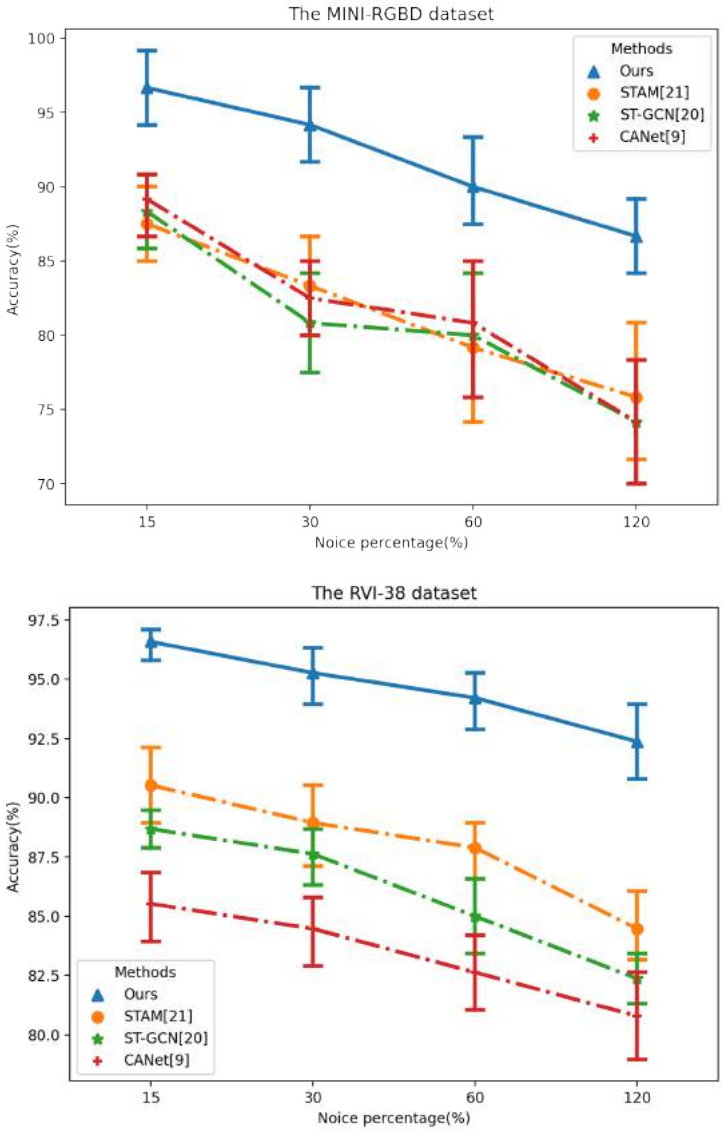


Figure 4.3: The robustness test compared with the state-of-the-art DNN methods. The short vertical bar of each method in different noise-level denotes the accuracy range between the first quartile and third quartile among all cross-validations. The line between each bar is linked by the mean accuracy value.

ing trend under different noise levels, which represents the stronger stability and robustness of our system.

4.4.9 Attention Analysis

Fig. 4.4 visualizes the interpretability of proposed attention module by presenting the attention value of each joint among all leave-one-out cross-validations on

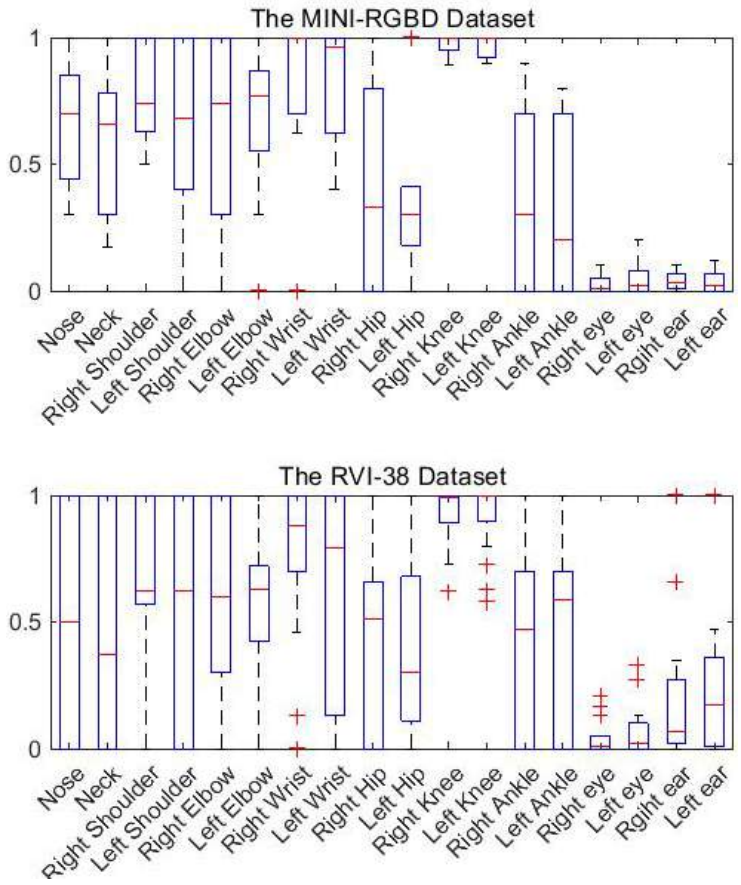


Figure 4.4: The visualization of attention weights of different joints among all cross-validations on each dataset.

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 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.

4.5 Discussions and Limitations

Clinical Significance. Our proposed system aims to provide undiagnosed CP patients with low-cost, non-intrusive CP abnormal movement classification results as a warning sign. This provides a way for supporting the early prediction of CP in the clinical resource-limited areas, and relieving the labour stress of expert-based GMA [15]. In addition, our system can provide clinicians with information about the importance-ranking of joints' movement and movement frequency in CP classification by visualizing the attention maps.

Limitation Discussions. During the research process of designing the data preprocessing pipeline, we observed that our system's performance depends on the accuracy of the estimated pose features. Specifically, since infants' movements generally involve twisting and rolling of the limbs and torso, when processing certain video frames with severe self-occlusion, OpenPose still extracts incorrect joint positions or positions with extremely low confidence scores. These low-quality pose features reduce the accuracy of the classification model. To address this noise, we have employed smoothing operations between frames; however, this cannot guarantee completely correct pose features. Therefore, improving the reliability of the pose estimation process by adapting more advanced systems such as [151, 148] is essential in future work. Additionally, while we have proposed a frequency-binning strategy to reduce high-frequency noise, it might risk losing informative subtle movements. This is because the inherent limitations of the binning strategy theoretically led to a reduction in the sharp changes of joint positions, but it cannot automatically distinguish whether these sharp changes are caused by noise or are inherent features of the movement itself. Although the parameter analysis for different binning algorithm experiment shows that reducing the percentage of high-frequency bands in infant movements is effective for CP prediction, we still believe that FAIGCN is only a preliminary but insightful CP prediction model

for analyzing infant movement frequency. Future work may require more capable modeling of frequency domain features and the introduction of modeling of spatial features, thereby enabling the FAIGCN model not only to maintain its efficacy in noise reduction but also to refine its sensitivity to the subtle dynamics of infant movements. Moreover, the limited size of the dataset contributes to the risk of overfitting during the network design process. An explainable and reliable data augmentation method is recommended for future studies to mitigate this issue.

4.6 Conclusion

In this chapter, 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 visualize the important joints that the network considers in CP prediction. The system is developed in PyTorch and fully released for future development. Experimental results show the importance of low-to-mid frequency data and the effectiveness and robustness of our system 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.

Statements and Declarations

This research is supported in part by the EPSRC NortHFutures project (ref: EP/X031012/1).

Chapter 5

Improving Interpretable Cerebral Palsy Prediction with Spatial-Frequency Analysis

5.1 Introduction

In our CP prediction study in Chapter 4, we highlighted significant differences in joint motion frequencies between CP infants and healthy groups, which can be effectively learned by frequency-informed GCNs. In addition, we further identified three limitations in existing CP prediction systems, which we propose to address in this chapter. First, there is still a lack of a holistic framework to learn both spatial and frequency information [26, 28], where each domain information could be complementary to the other domain information, contributing to a more comprehensive understanding of CP. This is because analysis in the frequency domain primarily focuses on identifying specific patterns, such as abnormal muscle contractions and rhythmic movements, whereas spatial domain analysis provides insights into the general physical movements and postures of individuals. Second, while interpretability is crucial for clinical validation and real-world application [23], existing work has limited focus on visualizing the decision process of complex deep

learning [20, 26]. For example, our system in Chapter 4 only interprets the CP in the frequency domain, but it lacks the explanation in the spatial-temporal domain. Third, whilst the efficacy of pose-based methods is intrinsically linked to the quality of pose features, the de facto pose extraction method, OpenPose, results in less accurate body joints particularly during self-occlusion [29, 26].

To address the above key challenges, in this chapter, we put forward a Spatial and Frequency Attention-based Graph Convolutional Networks (SAFA-GCN) to predict CP by fusing both spatial pose and movement frequency features. The spatial-frequency fusion design leverages the strengths of both domains and provides a holistic understanding of the underlying factors contributing to CP with enhanced capacity. It is inspired by the clinical GMA criteria [152] and clinical observation [24] about the CP infants' abnormal movement. These references emphasize the significance of categorizing abnormal movements in CP infants based on both posture appearance and movement fluency, which are respectively evident in the spatial and frequency domains. Furthermore, to improve both the interpretability and the prediction accuracy, we propose the spatial- and masked frequency-attention mechanisms, along with a clipping-and-fusion method to interpret the significance of which joints, frequency bands, and time slots contributes to the CP prediction, allowing the results to be verified by clinicians. Note that the masked frequency-attention mechanism is directly inspired and improved from the frequency-binning design in Chapter 4. Finally, we improve the accuracy and robustness of the predominated pose extraction process [28, 20, 22, 51, 26] by estimating pose features from RGB videos via AlphaPose [148], and conduct a dataset supplement to establish a performance benchmark for leading methods.

We validate the accuracy and robustness of our system by introducing a robust evaluation protocol with two data splitting strategies, namely leave-one-out cross-validation, 3-fold and 5-fold cross-validations. This new protocol particularly evaluates the system robustness when training with decreasing training set size in the same dataset. We compare our methods with the state-of-the-arts on the MINI-

RGBD dataset [27] and the RVI-38 dataset [28]. MINI-RGBD is widely used for CP classification performance comparison [29, 143, 144, 145]. RVI-38 is a comprehensive dataset collected by our team [28] for a more challenging CP prediction task, with a larger size of data captured during routine clinical care in daily situations. Experimental results demonstrate that our system outperforms state-of-the-art CP prediction methods on both datasets. Our interpretable system also visualizes the significance of joints, frequency bands, and time periods in the prediction, aiding clinicians in making more robust diagnostic decisions.

We open our code and the dataset supplement of the MINI-RGBD and RVI-38 datasets for validation and further development: `https://github.com/zhz95/SAFA-GCN` Comparing with our previous system in Chapter 4, our contributions are as follows:

- We design a novel two-stream architecture to fuse complementary spatial poses and movement frequency features for more robust and accurate CP prediction, validated with MINI-RGBD and RVI-38 using a more robust evaluation protocol.
- We improve both the interpretability and prediction accuracy of the system by proposing:
  (1) Spatial-wise and frequency-wise attention modules, allowing the visualizations of which joints and frequency bands contribute to a prediction.
  (2) A clipping-and-fusion method that analyzes individual temporal clips and fuses the results for a final prediction, allowing the visualization of when the movementis significant for making a prediction.
- We supplement the MINI-RGBD and RVI-38 datasets with new and more accurately extracted posture features, and conduct a comprehensive benchmark analysis on leading methods, demonstrating a consistent performance enhancement.

5.2 SAFA-GCN: A Spatial and Frequency Attention Based Graph Convolution Network

We use a CP prediction framework similar to Chapter 4, but the classification network is different. As shown in Figure 5.1, the input is RGB videos of the infants in a supine position, captured from a camera positioned above, as per the GMA guidelines [28]. Then, we extract 2D coordinates pose features by AlphaPose algorithm [148]. Each pose sequence is then divided into multiple clips with a sliding overlapped window and assign positional encoding to each clip. The pose features of each clip are passed to our novel Spatial and Frequency Attention based GCN (SAFA-GCN), which leverages spatial pose and movement frequency features for robust and interpretable CP prediction.

Our SAFA-GCN is a fully upgraded version of FAIGCN in Chapter 4. It not only provides more comprehensive interpretability and attention visualization, but also improves the accuracy of pose features and overall improves the robustness of the CP system. Specifically, the spatial stream adopts the 2D Euclidean joint position features based on a pose graph to model the spatial relationship between different joints. Meanwhile, the frequency stream employs the Fast Fourier Transform (FFT) on each joint's sequence to transform the spatial features to the frequency domain, enabling us to model the relationship between different frequency bands. In addition, the attention modules in each stream highlight and visualize the significance of individual joints and frequencies, supporting clinicians with more precise and robust decision-making. After that, the stream-level fusion module integrates the contribution of each stream by concatenating the latent representations of each stream. For the final CP prediction, a clip-level fusion is proposed to improve the prediction accuracy by integrating local temporal information from multiple clips and enhance the temporal interpretability by visualizing the significant time slot in CP prediction.

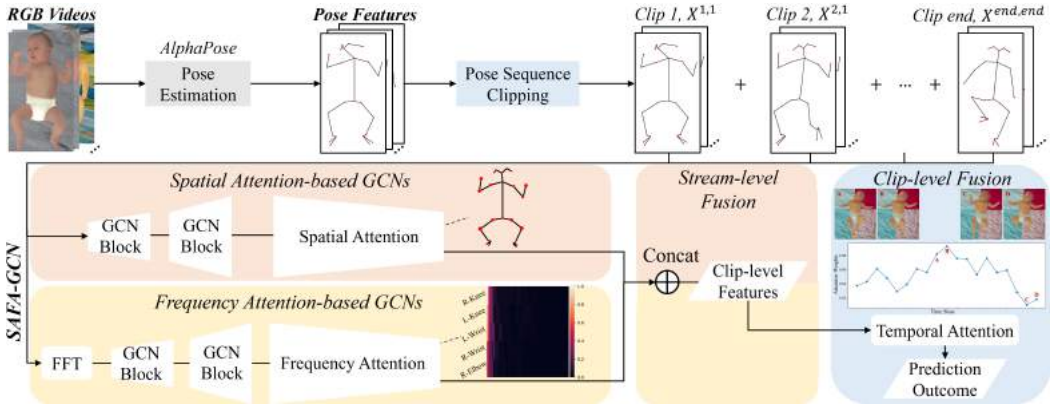


Figure 5.1: The overview of our proposed framework, where the input is the RGB videos of each individual. We employ AlphaPose to extract pose features, and obtain the pose sequence clips by pose sequence clipping. The clips of estimated pose sequences are fed into our two-stream CP prediction model, SAFA-GCN. Then, the SAFA-GCN fuses the prediction scores of each stream by the stream-level fusion module to obtain clip-level scores. The clip-level fusion module outputs the final prediction scores of each individual.

5.2.1 Spatial Attention Based GCNs

CP infants' pose usually suffer spatial impairment [153], specifically, they prefer to use only one side of their body with the presence of the fidgety movement (FMs). FMs refer to the infant joint movement (i.e., neck, trunk and limbs) in small amplitude and moderate speed with varying accelerations and directions [15]. Therefore, we introduce a spatial attention-based GCN stream to model the infant's pose information in the form of joint positions. Compared to our CP prediction network in Chapter 4, which only used frequency-attention, such a spatial-attention stream significantly improves the interpretability of the system by directly modeling the infants' spatial movements.

Spatial GCNs As shown in Fig. 5.2, each spatial GCN includes a GCN layer, an MLP, a batch normalization layer, a dropout layer and an ReLu activation layer. The input feature of the spatial stream is the set of the clipped 3D video sequences $\{X^{d,i}|d \in D, i \in \frac{N}{C}\}, X^{d,i} \in \mathbb{R}^{T\times C\times J}$, where D is the total number of clips in video i, N is the batch size, $\frac{N}{C}$ is the number of videos in the batch, T is the length

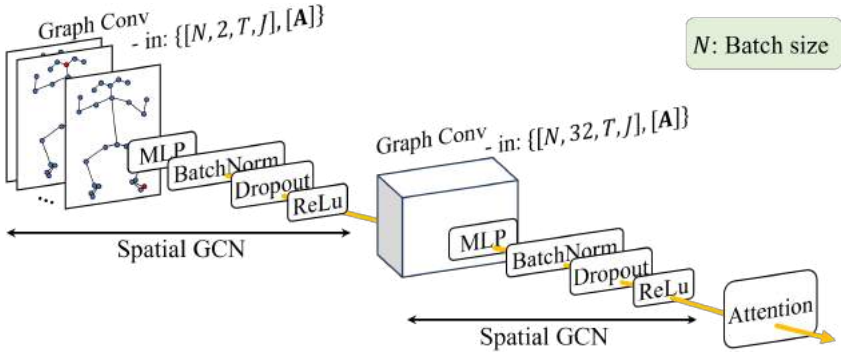


Figure 5.2: The architecture detail of spatial attention-based GCNs.

of frames, C is the 2D Euclidean joint position estimated by AlphaPose, and J is the total number of the joints. We apply a spatial pose graph $G = (V, E)$ aligned with the human skeletal graph to structure spatial domain pose data. In this graph, $V = \{v_{p,q} | p = 1, ..., T; q = 1, ..., J\}$ denotes the spatial domain node set represented by the joints positions, and $v_{p,q}$ is the position of q-th joint at p-th frame. The spatial edge set E depicts the natural connections of human joints at each frame. The corresponding graph convolutional operation [121] is the same as the Eq.(4.3) in Chapter 4.

Instead of employing the temporal convolution, we adopt the Multi-layer Perceptron (MLP) to model the inter-frame features and aggregate the frame-level representations to obtain spatial stream representations $\hat{y}_s$, since the MLP is suitable for our objective of learning global information related to FMs across feature pairs [154] within a short 100 frames window.

The Spatial Attention Mechanism We propose to apply the following attention mechanism for more informative and discriminative global representation learning by its characteristics on adaptive feature selection. It also enhances the CP prediction interpretability by quantifying the relative importance of human joints. To achieve this, we aggregate the spatial representations at frame p obtained from spatial GCN $\{\mathbf{h}_{p,1}, \mathbf{h}_{p,2}, \ldots, \mathbf{h}_{p,J}\}$ with attentions $\alpha_{p,q}$, by Eq.(5.1) to obtain the

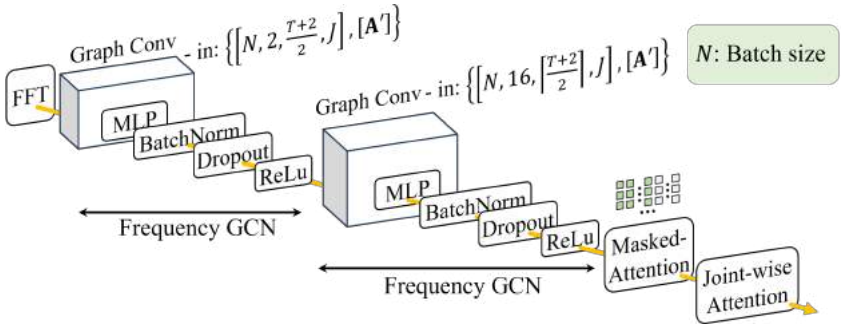


Figure 5.3: The architecture detail of frequency attention-based GCNs.

frame-level representation $\mathbf{s}_p$:

$$\mathbf{s}_p = \sum_{q=1}^{Q} \alpha_{p,q} \mathbf{h}_{p,q}, \tag{5.1}$$

in which the spatial attention weight $\alpha_{p,q}$ is defined as:

$$\alpha_{p,q} = \frac{\exp\left(\mathbf{w}_\alpha^\mathsf{T} tanh\left(\mathbf{W_h} \mathbf{h}_{p,q}\right)\right)}{\sum_p \exp\left(\mathbf{w}_\alpha^\mathsf{T} tanh\left(\mathbf{W_h} \mathbf{h}_{p,q}\right)\right)}, \tag{5.2}$$

where $\mathbf{W_h}$, $\mathbf{w}_\alpha$ are learnable parameters.

5.2.2 Frequency Attention-based GCNs

We propose a second stream to model the CP prediction in the frequency domain. It is driven by our research insights of modeling frequency features with filtering high-frequency noise in Chapter 4, and the clinical observation that highlights the disparity in body movement frequencies between healthy infants and those who have experienced CP [24]. Integrating the frequency stream sensitive to movement regularity, pace, and smoothness provides a comprehensive supplement to coordination and appearance information from the single spatial stream. Existing deep learning-based CP prediction networks [49, 51, 22] other than our prior study [26] lack the important analysis of frequency-domain information [24], thus limiting their performance and system interpretability. In Chapter 4, we introduced a frequency-binning strategy that compresses frequency information into different bins. However, that method results in some information loss in mid-to-high frequency bands and reduces the system's robustness. This limitation becomes

particularly pronounced when modeling more diverse data distributions. To this end, we propose a novel approach: a masking frequency attention mechanism to restrict the attention to non-high frequency noise while preserving all frequency information.

Frequency GCNs Similar to the frequency feature transform in Chapter 4, as shown in Fig. 5.3, the frequency stream first obtains the frequency features by transforming each joint's 2D position sequence x_n into the frequency domain via Fast Fourier Transform (FFT) [147]:

$$X_k = \sum_{n=0}^{A-1} x_n e^{\frac{-i2\pi bn}{A}}, \quad b = 0, \ldots, N-1, \tag{5.3}$$

where $e^{\frac{i2\pi}{A}}$ is a primitive A^{th} root of 1.

In frequency stream, we propose a frequency domain chain graph $G' = (V', E')$, where $V' = \{v'_{q,c} | q = 1, ..., J; c = 1, ..., \lceil \frac{T+2}{2} \rceil\}$ denotes the frequency domain node set represented by the FFT coefficients, and $v'_{q,c}$ represents the vertex of the c-th frequency band of q-th joint, note that the size of the frequency band is reduced to $\lceil \frac{T+2}{2} \rceil$ due to the FFT symmetry. The edge set is defined by the inter-frequency edge set $E' = \{v'_{q,c} v'_{q,c+1}\}$.

The Frequency Attention Mechanism We propose a masked attention to learn the weight of different frequency bands while masking the high-frequency representations. The aggregation function of the frequency representations

$\{\mathbf{f}_{q,1}, \mathbf{f}_{q,2}, \ldots, \mathbf{f}_{q,\lceil \frac{T+2}{2} \rceil}\}$ of joint q with attention $\beta_{q,c}$ is formulated as:

$$\mathbf{u}_q = \sum_{c=1}^{C} \beta_{q,c} \mathbf{f}_{q,c}, \tag{5.4}$$

in which the frequency attention weights $\beta_{q,c}$ is defined as:

$$\beta_{q,c} = \frac{\exp\left((\mathbf{w}_\beta^\mathsf{T} + \mathcal{M}) tanh\left(\mathbf{W}_\mathbf{f} \mathbf{f}_{q,c}\right)\right)}{\sum_j \exp\left((\mathbf{w}_\beta^\mathsf{T} + \mathcal{M}) tanh\left(\mathbf{W}_\mathbf{f} \mathbf{f}_{q,c}\right)\right)}, \tag{5.5}$$

where $\mathbf{W_f}$, $\mathbf{w}_\beta$ are learnable parameters, and the attention mask $\mathcal{M}$ at location (q, c) is:

$$\mathcal{M}(q,c) = \begin{cases} 0 & \text{if} \quad c < r\lceil\frac{T+2}{2}\rceil \\ -\infty & \text{otherwise} \end{cases}. \tag{5.6}$$

Motivated by the clinical observation [24] and pilot study [26], the attention mask $\mathcal{M}$ is designed to modulate the weights of features at the frequency bands below or above $r\lceil\frac{T+2}{2}\rceil$, where r is a constant ratio to define the high-frequency band.

We further aggregate the joint q's frequency aggregation representation $\mathbf{u}_q$ to quantify the joint-wise importance γ_q and obtain the frequency stream prediction score $\hat{y}_f$. This enables validating the frequency stream attention by intuitively comparing the joint-wise importance with the spatial stream attention, such that improves the system's robustness.

$$\hat{y}_f = \sum_{q=1}^{Q} \gamma_q \mathbf{u}_q, \tag{5.7}$$

$$\gamma_q = \frac{\exp\left(\mathbf{w}_\gamma^\intercal tanh\left(\mathbf{W_u}\mathbf{u}_q\right)\right)}{\sum_q \exp\left(\mathbf{w}_\gamma^\intercal tanh\left(\mathbf{W_u}\mathbf{u}_q\right)\right)}, \tag{5.8}$$

where $\mathbf{W_u}$, $\mathbf{w}_\gamma$ are learnable parameters.

5.2.3 Stream-Level Fusion

We apply stream-level late fusion to fuse scores from both the frequency and spatial domains, which provides better prediction robustness as errors from two models are dealt with independently [155].

To preserve the information from different streams, we fuse features from both streams by concatenation and obtain the clip-level representations:

$$\hat{y}_l = concat(\hat{y}_s, \hat{y}_f), \tag{5.9}$$

We prefer the late fusion architecture to the early fusion [22], since the latter may force the model to learn redundant or irrelevant features in processing

the heterogeneous frequency and spatial data [155], and constrain the model interpretability in multiple domains. The rationale of our two-stream design is based on the invariance between spatial and frequency domains from signal processing theory [156].

5.2.4 Clipping-and-Fusion

We introduce a clipping-and-fusion method to improve system prediction performance and temporal domain interpretability that was lacking in Chapter 4. Pose sequence clipping enables better capturing of critical local temporal features during CP prediction, which aligns with the clinical CP diagnosis of observing the FMs patterns in GMA [15]. Clip-level fusion fuses individual clip results for the final prediction, with an attention mechanism for visualizing clip importance.

Pose Sequence Clipping Despite working with relatively smaller datasets, most existing CP prediction methods [29, 143, 28, 20, 26] use entire video sequences as input. This approach presents a challenge as their models face the risk of overfitting and may lack robustness when the train-test split is altered. A more effective strategy is to clip the videos into small segments, which allows the model to better capture crucial local features and increase the amount of data to reduce the risk of overfitting [157].

To facilitate more robust CP predictions, we propose a pose sequence clipping process to clip each 3D video pose sequence $X \in \mathbb{R}^{T\times C\times J}$ into multiple 100 frames (3.3s) clips with a sliding overlapping window size of 50 frames, denoted by $X^i = [X^{1,i}, ...X^{d,i}, ..., X^{end,i}]$, where $X^{d,i}$ is the d-th 100-frame clip of the video i. The overlapping window size value chosen to be half of the clip length (50 frames) is considered by its temporal interpretability while reducing the impact of duplicated information. In addition, to ensure that clips belonging to the same video are not split into different training/testing sets, we apply positional encoding based on the original temporal order to all clips originating from the same video.

Clip-Level Fusion We propose a clip-level fusion to integrate the local temporal information from multiple clips of the same individual for a comprehensive global temporal representation. Compared with learning the temporal information from the entire video sequence [28, 49, 26], our design better captures key local features that may be overlooked in the full sequence, which is also consistent with the clinical CP prediction of observing the FM patterns in GMA [15].

We propose a weighted average based aggregation to fuse the clip-level representations and make this process interpretable by introducing a temporal attention mechanism:

$$\hat{y} = Softmax(\sum_{l=1}^{L} \lambda_l \hat{y}_l), \tag{5.10}$$

$$\lambda_l = \frac{\exp\left(\mathbf{w}_\lambda^\intercal tanh\left(\mathbf{W}_{\hat{y}} \hat{y}_l\right)\right)}{\sum_l \exp\left(\mathbf{w}_\lambda^\intercal tanh\left(\mathbf{W}_{\hat{y}} \hat{y}_l\right)\right)}, \tag{5.11}$$

where $\mathbf{W}_{\hat{y}}$, $\mathbf{w}_\lambda$ are learnable parameters.

5.3 The Data Processing Pipeline

We propose to employ an empirically more precise pose estimation method [148] to enhance CP prediction accuracy and robustness. This is motivated by the observation from Chapter 4, in which the performance of existing automated CP prediction systems relies heavily on the quality of input pose features, and inaccurate pose features bring significant noise to the system.

5.3.1 Improved Feature Extraction for More Accurate Poses

In Chapter 4, we extract 2D coordinates pose features from videos for CP prediction. Whist the pose features bring remarkable interpretability and robustness, and avoid the background noise compared to using the image features directly, we observed that performance of FAIGCN depends on the accuracy of the estimated pose features by the de facto OpenPose pose estimation [1]. Similar arguments

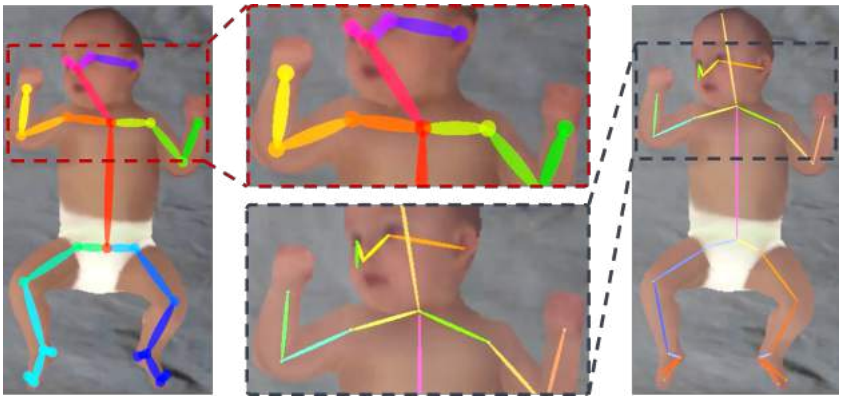


Figure 5.4: The comparison between OpenPose (left) and AlphaPose (right) on a sample frame. Note AlphaPose's higher neck and hips qualities.

are also proposed in [20, 28, 51, 22]. Specifically, since infants' movements generally involve twisting and rolling of the limbs and torso, when processing certain video frames with severe self-occlusion, OpenPose still extracts incorrect joint positions or positions with extremely low confidence scores. These low-quality pose features reduce the accuracy of the classification model. One possbile reason is that OpenPose has not theoretically minimised the quantization errors from its heatmap representations calculation [158].

To improve the precise of extracted pose features for more accurate CP prediction, we employ a state-of-the-art pose estimation algorithm, AlphaPose [148], to extract more accurate 2D skeleton features. This module mimics the motion observing step of GMA assessors. As shown in Fig. 5.4, the AlphaPose joint positions of shoulder-neck-hip area are significantly more precise than those of OpenPose. In addition, AlphaPose successfully detects the left-ear keypoint while OpenPose fails due to self-occlusion. Moreover, we keep the data pre-processing pipelines in Chapter 4 to improve the feature quality and mitigate the overfitting risk.

5.4 Experiments

In this section, we evaluate the performance of our method and the other ten models on two datasets by involving a robust evaluation protocol. We also conduct

multiple experiments to validate the robustness and interpretability of our system.

5.4.1 Implementation Details

For spatial stream, we use two GCN blocks with two convolutional layers with 32, 64 output channels respectively. For frequency stream, the output channel sizes of two corresponding convolutional layers are 16, 32. We adapt the focal-loss [159] to tackle the class imbalance. The optimizer is *Adams*, and the training hyperparameters are: (i) MINI-RGBD dataset, *batch size* = 2, *learning rate* = 0.0001 with 0.1 decay every 100 epoches, *Max Epoch* = 500; (ii) RVI-38 dataset, *batch size* = 4, *learning rate* = 0.001 with 0.1 decay every 100 epoches, *Max Epoch* = 500.

5.4.2 Computational Environment

We conducted the experiments on an Ubuntu 18.04 PC equipped with an NVIDIA GeForce RTX 3080. The total model training time of SAFA-GCN on MINI-RGBD is about 1.5 hour, encompassing the computation of the joints' position from RGB videos. The average time cost for CP prediction on a video of 1000 frames (approximately 33 seconds) is around 70 seconds. This suggests our system's potential to be employed in real-time interactive diagnosis.

5.4.3 Dataset

We verify our models on the MINI-RGBD and the RVI-38 dataset, which are described in Chapter 4.

5.4.4 Evaluations with a Robust Evaluation Protocol

To evaluate the effectiveness, robustness, and reliability of our proposed system comprehensively, we propose a robust evaluation protocol by conducting different

types of cross-validation settings among two datasets, and present a benchmark study based on this new evaluation. This is motivated by our finding that most existing methods [29, 28, 21, 20, 26] are negatively affected by the decreasing training set size, which means that the existing leave-one-out cross-validation (LOOCV) based evaluation [29, 28, 21, 20, 26, 51] is not sufficient for validating the system robustness.

We employ leave-one-out cross-validation (LOOCV), 5-fold CV, and 3-fold CV for the RVI-38 dataset, and LOOCV and 3-fold CV for the MINI-RGBD dataset due to data size constraints. The count of CP class in the training set decreases as the size of each fold increases, posing a challenge for model training as it increases the overfitting risk in a smaller, imbalanced training set. We further propose an additional robustness analysis by adding different levels of Gaussian noise to the pose features for evaluating the model's sensitivity and resilience to perturbations, ensuring its reliability in real-world scenarios with potential noise and variability.

In order to verify the effectiveness and robustness of our proposed method, we compare SAFA-GCN with seven state-of-the-art DNN-based CP prediction methods FCNet [29], Conv1D [29], Conv2D [29], CANet [20], STAM [22], FAIGCN [26], and WO-GMA [51], as well as machine learning based methods Ens-1 [143], Ens-2 [28] with different hand-crafted features. We also follow [22, 26] to compare with ST-GCN [125] as it is an effective GCN for human pose data classification task. To the best of our knowledge, we compare our method with major existing vision-based computer-aided CP prediction methods. Our evaluation uses four metrics: prediction accuracy (AC), sensitivity (SE), specificity (SP), and the F1-Score (F1).

Method	Features	MINI-RGBD 3-fold CV AC	SE	SP	F1	RVI-38 5-fold CV AC	SE	SP	F1	RVI-38 3-fold CV AC	SE	SP	F1
Our SAFA-GCN	Pose + Frequency	**100.00**	**100.00**	**100.00**	**100.00**	**97.37**	**83.33**	**100.00**	**90.91**	**97.37**	**83.33**	**100.00**	**90.91**
FCNet [29]*	HOJO/D2D	83.33	75.00	87.50	75.00	-	-	-	-	-	-	-	-
Conv1D [29]*	HOJO/D2D	83.33	75.00	87.50	75.00	-	-	-	-	-	-	-	-
Conv2D [29]	Pose	91.67	**100.00**	87.50	88.89	78.94	16.67	90.63	20.00	76.32	16.67	87.50	18.18
CANet [20]	Pose	91.67	**100.00**	87.50	88.89	84.21	33.33	96.77	44.44	81.58	16.67	93.75	22.22
ST-GCN [125]	Pose	91.67	**100.00**	87.50	88.89	86.84	66.67	90.63	61.54	86.84	66.67	90.63	61.54
STAM [22]	Pose	91.67	**100.00**	87.50	88.89	92.11	**83.33**	93.75	76.92	89.47	66.67	93.75	62.50
FAIGCN [26]	Frequency	91.67	**100.00**	87.50	88.89	94.74	**83.33**	96.88	83.33	94.74	**83.33**	96.88	83.33
WO-GMA [51]	Pose	91.67	**100.00**	87.50	88.89	94.73	**83.33**	96.88	83.33	94.74	**83.33**	96.88	83.33
Ens-1 [143]*	HOJO+D2D	91.67	**100.00**	87.50	88.89	94.74	**83.33**	96.88	83.33	92.11	**83.33**	93.75	76.92
Ens-2 [28]*	Velocity+Pose	**100.00**	**100.00**	**100.00**	**100.00**	94.74	**83.33**	96.88	83.33	92.11	**83.33**	93.75	76.92

* denotes the methods that use the hand-crafted features.

Table 5.1: The 5-fold and 3-fold cross-validations comparisons with the state-of-the-arts on two datasets

5.4.5 Performance Test

We compare CP prediction performance of our SAFA-GCN with the state-of-the-art methods in Table 5.1 and 5.2. We obtain the following observations: (1) Comparing with the proposed FAIGCN in Chapter 4, our SAFA-GCN consistently exhibits superior performance as its AC, SE, SP, and F1 achieve the highest score across all data-splittings on two datasets, showcasing the robustness and reliability of our new system. (2) Comparing the conventional evaluation protocol results in Table 5.1, Table 5.2 shows that our new evaluation protocol effectively verifies the performance and robustness of different methods and brings challenges to model training. Specifically, We have identified a noticeable decrease in the metrics of other methods, especially the AC, SP and F1, as we transition from LOOCV to 5-fold or 3-fold CV. This decline is largely ascribed to the class imbalance in both datasets, wherein the CP class is underrepresented compared to the healthy class. Such a result supports the motivation of our proposed new evaluation protocol. In contrast, our method consistently preserves its performance. (3) We observe the advantages of graph structure modeling for human pose features as ST-GCN [125], STAM [22], FAIGCN [26], WO-GMA [51] and SAFA-GCN outperform other methods that rely on CNNs. (4) In addition to WO-GMA [51] which adopts weakly supervised learning, the superiority of SAFA-GCN and FAIGCN over other supervised learning methods reflects the advantages of considering frequency analysis in CP prediction. (5) An interesting finding is that the machine learning-based method, Ens-2 [21], which utilizes an ensemble of classification models (i.e., SVM, LR, LDA, and DT) on the fused hand-crafted features [143, 21], shows very close performance to our SAFA-GCN on both datasets, except for the 5-fold and 3-fold CV in RVI-38. On the one hand, this observation demonstrates the superiority of handcrafted features in classification tasks; on the other hand, it highlights the effectiveness and generalizability of our SAFA-GCN, which requires only raw pose features without necessitating complex, manually-defined feature engineering.

	MINI-RGBD				RVI-38			
Method	AC	SE	SP	F1	AC	SE	SP	F1
SAFA-GCN	**100.00**	**100.00**	**100.00**	**100.00**	**97.37**	**83.33**	**100.00**	**90.91**
FCNet [29]	91.67	**100.00**	87.50	88.89	-	-	-	-
Conv1D [29]	83.33	75.00	87.50	75.00	-	-	-	-
Conv2D [29]	91.67	**100.00**	87.50	88.89	81.58	33.33	90.63	36.36
CANet [20]	91.67	**100.00**	87.50	88.89	86.84	66.67	90.63	61.54
ST-GCN [125]	91.67	**100.00**	87.50	88.89	89.47	66.67	93.75	62.50
STAM [22]	91.67	**100.00**	87.50	88.89	92.11	**83.33**	93.75	76.92
FAIGCN [26]	**100.00**	**100.00**	**100.00**	**100.00**	**97.37**	**83.33**	**100.00**	**90.91**
WO-GMA [51]	**100.00**	**100.00**	**100.00**	**100.00**	**97.37**	**83.33**	**100.00**	**90.91**
Ens-1 [143]	91.67	**100.00**	87.50	88.89	94.74	**83.33**	96.88	83.33
Ens-2 [28]	**100.00**	**100.00**	**100.00**	**100.00**	**97.37**	**83.33**	**100.00**	**90.91**

Table 5.2: The LOOCV comparisons with the state-of-the-arts

5.4.6 Additional Robustness Analysis

To demonstrate that our proposed method helps improve the robustness, we make the CP prediction task more challenging by adding different levels of Gaussian noise to the estimated pose features (Fig. 5.5). The noise levels are divided into different levels, ranging from 15% to 180% of the standard deviation of the raw pose sequence. To visualize a more pronounced performance change in a larger dataset, we utilize larger noise level ranges for RVI-38 dataset compared with MINI-RGBD dataset. The average accuracy is reported among ten different training seeds based on the LOOCV data-splitting strategy. 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, the experimental evidence reveals that SAFA-GCN suffers from the least performance drop as the noise level increases on both datasets, reflecting the stronger robustness of our proposed system.

5.4.7 Model Interpretability

In this section, we make the decision of our system be able to justified by clinicians to improve the reliability of the computer-aid decision-making. Motivated by the clinical interests in understanding the CP prediction [52, 24, 29], we visualize

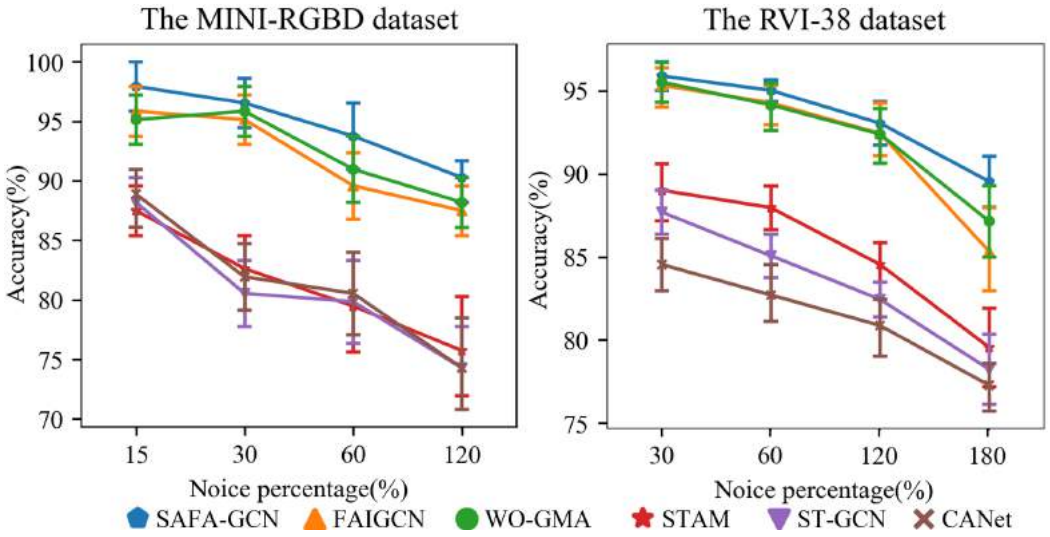


Figure 5.5: The robustness analysis compared with the state-of-the-art DNN methods. The short vertical bar of each method in different noise-level denotes the accuracy range between the first quartile and third quartile among all LOOCV cross-validations. The line between each bar is linked by the mean accuracy value.

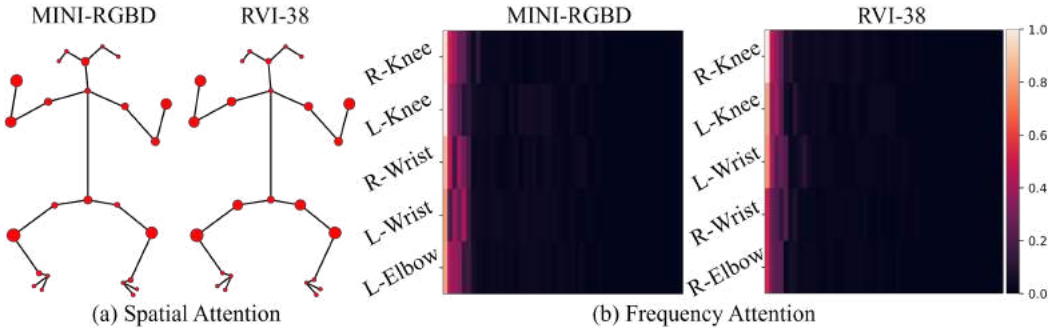


Figure 5.6: (a) The visualization of spatial attention weights of different joints among all cross-validations; (b)The frequency attention map of top-5 joints with highest joint-wise importance. The x-axis represents the spectrum, with frequencies increasing from left to right.

the important infants joints, joint movements frequency bands and time slots by our attention design.

Spatial Domain Interpretation Fig. 5.6 (a) visualizes the attention weights of each joint among all cross-validations on each dataset. We observe that in both datasets, the attention weight assigned to knees, wrists, and elbows significantly exceeds that of other parts. This shows that our system emphasizes monitoring the movements of infants' knees and wrists, which aligns with the fact that these

joints exhibit the most pronounced movements and frequency changes in the video recordings. Additionally, the attention weights for the eyes, ears, and heels are notably lower than those for others. This may be attributed to self-occlusion issues that introduce noise into the AlphaPose estimations of these joints.

Frequency Domain Interpretation Fig. 5.6 (b) visualizes the normalized averaging attention value of the top-5 joints with the highest frequency stream spatial attention on each dataset. We have observed that attention is primarily concentrated in the low-frequency range (i.e., the first 8 frequency bands), and the highest attention values for each joint occur in the first frequency band. Considering the physical meaning of the first frequency band, it represents the global offset or bias of the original pose sequence. Therefore, this finding inspires us to consider that the global offset or bias in spatial-temporal sequences may be a feature with predictive capabilities for CP prediction. Additionally, the concentration of attention in the low-frequency range aligns with our hypothesis that high frequencies provide minimal information. This result further supports the feasibility of using frequency masking method.

Temporal Domain Interpretation Fig. 5.7 visualizes the temporal attention weights and the corresponding infant frames in a MINI-RGBD video, with each time slot referring to a clip. We find that infants' movement amplitude and speed are generally smaller in time slots (e.g., A and B) with high attention weight. In contrast, infant presents pronounced limb movements in the time slots (e.g., C and D) with low attention weights. This result is consistent with the GMA criteria [152], showing the reliability of our proposed method.

5.4.8 Component Analysis and Ablation Studies

Comparison with Different Clip-Level Fusion Methods: Motivated by the fusion strategy evaluation in [160], we compare our non-linear MLP-based temporal

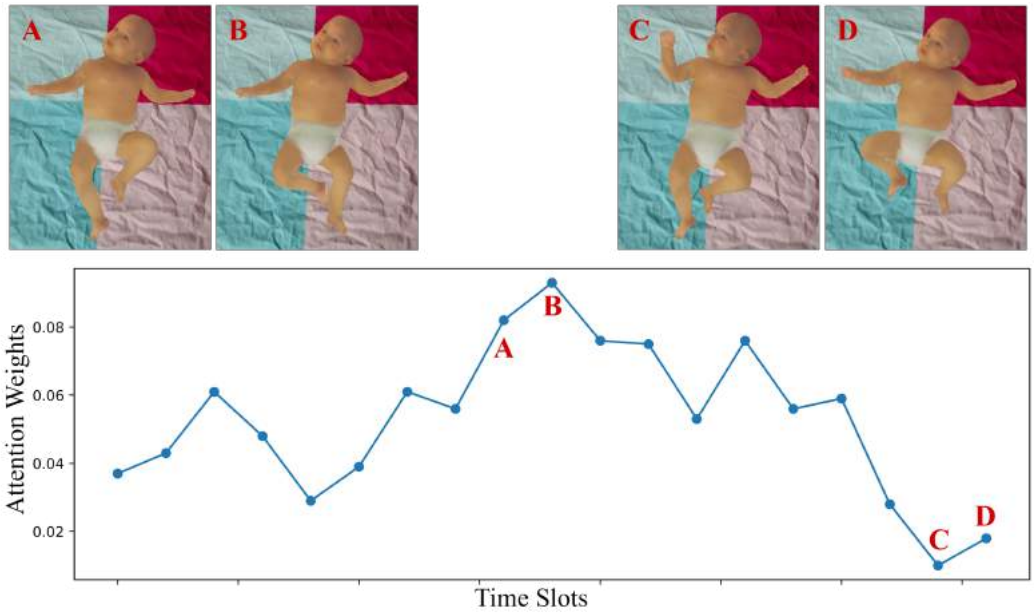


Figure 5.7: A visualization example of temporal attention weights in a MINI-RGBD video. The index letters A-D map the infant images to the corresponding time periods and attention weights.

attention fusion with the linear fusion method using a hard margin linear SVM classifier. This comparison provides a robust evaluation, considering scenarios where the boundaries of multiple clip representations are linear [161]. From Table 5.3, we observe that the performance of our fusion method achieves the best performance across all metrics in all CVs. In addition, fusing by training a linear SVM model yields inferior results compared to ours, which is pronounced in the 3-fold CV data-splitting approach. This can be explained by the different distribution of data in the 3-fold CV setting, emphasizing the adaptability of our MLP-based temporal attention fusion method across various data distribution scenarios.

	Method	AC	SE	SP	F1
LOOCV	(i) Late fusion				
	SAFA-GCN	97.37	83.33	100.00	90.91
	Clip-level SVM fusion	97.37	83.33	100.00	90.91
	(ii) Early-fusion	97.37	83.33	100.00	90.91
5-fold CV	(i) Late fusion				
	SAFA-GCN	97.37	83.33	100.00	90.91
	Clip-level SVM fusion	97.37	83.33	100.00	90.91
	(ii) Early-fusion	97.37	83.33	100.00	90.91
3-fold CV	(i) Late fusion				
	SAFA-GCN	97.37	83.33	100.00	90.91
	Clip-level SVM fusion	94.74↓	83.33	96.88↓	83.33↓
	(ii) Early-fusion	94.74↓	83.33	96.88↓	83.33↓

Table 5.3: Comparisons of Different Fusion Designs on the RVI-38

Comparing the Late-fusion with Early-fusion: We further investigate the effectiveness of the early-fusion method instead of the late-fusion, as it is an alternative common fusion method to fuse the information from two streams [28, 49]. We implement an early-fusion system that fuses the pose features and FFT features by concatenation. Then, we feed the fused features to our spatial attention-based GCNs and generate the CP prediction score by the clip-fusion. The 3-fold CV results in Table 5.3 show that our proposed late-fusion architecture have stronger robustness than concatenating-based early-fusion method, demonstrating the effectiveness of developing different GCNs to model different domain features.

Effects on Pose Estimation Algorithms: We compare pose features from AlphaPose with OpenPose in Table 5.4. AlphaPose features exhibit a remarkable improvement of approximately 3% across all evaluated methods when compared with OpenPose features. It demonstrates the efficacy of the AlphaPose algorithm in pose feature quality improvements.

Ablation Study: An ablation is conducted to evaluate the effectiveness of the proposed two-stream architecture, attention, and the clipping-and-fusion method, as shown in Table 5.5.

	Method	AC	SE	SP	F1
LOOCV	**OpenPose**				
	SAFA-GCN	97.37	83.33	100.00	90.91
	FAIGCN [26]	97.37	83.33	100.00	90.91
	STAM [22]	92.11	83.33	93.75	76.92
	Conv2D [29]	81.58	33.33	90.63	36.36
	CANet [20]	86.84	66.67	90.63	61.54
	AlphaPose				
	SAFA-GCN	97.37	83.33	100.00	90.91
	FAIGCN [26]	97.37	83.33	100.00	90.91
	STAM [22]	94.74↑	83.33	96.88↑	83.33↑
	Conv2D [29]	84.21↑	33.33	96.77↑	44.44↑
	CANet [20]	89.47↑	66.67	93.75↑	62.50↑
5-fold CV	**OpenPose**				
	SAFA-GCN	97.37	83.33	100.00	90.91
	FAIGCN [26]	94.74	83.33	96.88	83.33
	STAM [22]	92.11	83.33	93.75	76.92
	Conv2D [29]	78.94	16.67	90.63	20.00
	CANet [20]	84.21	33.33	96.77	44.44
	AlphaPose				
	SAFA-GCN	97.37	83.33	100.00	90.91
	FAIGCN [26]	94.74	83.33	96.88	83.33
	STAM [22]	92.11	83.33	93.75	76.92
	Conv2D [29]	84.21↑	33.33↑	96.77↑	44.44↑
	CANet [20]	86.84↑	66.67↑	90.63	61.54↑
3-fold CV	**OpenPose**				
	SAFA-GCN	94.74	83.33	96.88	83.33
	FAIGCN [26]	92.11	83.33	93.75	76.92
	STAM [22]	89.47	66.67	93.75	62.50
	Conv2D [29]	76.32	16.67	87.50	18.18
	CANet [20]	81.58	16.67	93.75	22.22
	AlphaPose				
	SAFA-GCN	97.37↑	83.33	100.00↑	90.91↑
	FAIGCN [26]	94.74↑	83.33	96.88↑	83.33↑
	STAM [22]	92.11↑	83.33↑	93.75	76.92↑
	Conv2D [29]	81.58↑	16.67	93.75↑	22.22↑
	CANet [20]	86.84↑	66.67↑	90.63	61.54↑

Table 5.4: Comparing the AlphaPose with OpenPose on the RVI-38

(i) **Effects of the attention mechanism:** We compare our full model with a variant without both attentions and taking the means instead, a variant with only frequency attention and calculates the spatial stream representations by averaging, and a variant with only spatial attention and calculates the frequency stream representations by averaging. Removing each type of attention mechanism leads to a significant drop in all metrics, especially the AC, SP, and F1. It shows the effectiveness of the proposed attention mechanism.

(ii) **Effects of the frequency attention mask:** We compare SAFA-GCN and its variant without the mask. The outcome shows that the frequency attention mask enhances performance and robustness since the full model outperforms the variant in 3-fold CV and achieves comparable results in LOOCV.

(iii) **Effects of the stream-fusion:** To validate whether any adverse effect exists on the two-stream fusion, we compare SAFA-GCN with its single-stream variants. It shows that the variants suffer a performance drop in some data-splitting methods, demonstrating that the complementary information provided by both streams effectively enhances the predictive capabilities and robustness.

(iv) **Effects of the clipping-and-fusion:** We compare the SAFA-GCN with a variant that utilizes solely the raw pose sequence as input, without incorporating the clipping-and-fusion process. Although the performance remains the same in LOOCV, it drops in the 3-fold CV in both datasets, indicating the efficacy of the clipping-and-fusion method.

Method	LOOCV				3-fold CV			
	AC	SE	SP	F1	AC	SE	SP	F1
Ours-full	**97.37**	**83.33**	**100.00**	**90.91**	**97.37**	**83.33**	**100.00**	**90.91**
Attention								
w/o both	94.74	**83.33**	96.88	83.33	89.47	66.67	93.75	62.50
w/o spatial	94.74	**83.33**	96.88	83.33	92.11	**83.33**	93.75	76.92
w/o frequency	94.74	**83.33**	96.88	83.33	89.47	66.67	93.75	62.50
Mask								
w/o	**97.37**	**83.33**	**100.00**	**90.91**	94.74	**83.33**	96.88	83.33
Stream-fusion								
w/o spatial	**97.37**	**83.33**	**100.00**	**90.91**	94.74	**83.33**	96.88	83.33
w/o frequency	94.74	**83.33**	96.88	83.33	94.74	**83.33**	96.88	83.33
Clippling-and-fusion								
w/o	**97.37**	**83.33**	**100.00**	**90.91**	94.74	**83.33**	96.88	83.33

Table 5.5: The ablation study on the RVI-38

Parameter Analysis of Frequency Mask Table 5.6 shows the parameter analysis of frequency mask ratio. Although the frequency attention (seen in 5.4.7) is dominated by the approximately first 15.7% of the frequency band, the mask ratio of 1/3 leads to a drop compared with $r = 1/2$ or $2/3$ in 3-fold CV. This suggests that masking excessive features may hinder the system's ability to learn attention distributions. Whilst the ratios of 2/3 and 1/2 provide the same performance on all data-splittings, we keep $r = 2/3$ in our final model since it provides more features for investigating the CP.

r	LOOCV				3-fold CV			
	AC	SE	SP	F1	AC	SE	SP	F1
1	**97.37**	**83.33**	**100.00**	**90.91**	94.74	**83.33**	96.88	83.33
2/3	**97.37**	**83.33**	**100.00**	**90.91**	**97.37**	**83.33**	**100.00**	**90.91**
1/2	**97.37**	**83.33**	**100.00**	**90.91**	**97.37**	**83.33**	**100.00**	**90.91**
1/3	**97.37**	**83.33**	**100.00**	**90.91**	94.74	**83.33**	96.88	83.33

Table 5.6: Comparison with frequency mask ratios on the RVI-38

5.5 Conclusion

In this Chapter, we propose two-stream attention based GCNs to automatically predict CP from RGB videos. We propose to fuse the complementary spatial and frequency features motivated by clinical observations for reliable CP prediction.

We propose the spatial attention, frequency attention, and a clipping-and-fusion method to strengthen the prediction reliability and interpretability. Our system visualizes important human joints, frequency bands and time ranges in CP prediction to support clinicians in making accurate and robust decisions. We also supplement two datasets with more accurate posture features and provide a performance benchmark analysis of leading methods. In the future, focusing on the challenge of collection medical data, we will use generative models (e.g., diffusion models [162], VAE [163]) to obtain a large synthetic CP dataset.

5.6 Statements and Declarations

This research is supported in part by the EPSRC NortHFutures project (ref: EP/X031012/1).

Chapter 6

Pose-Based Tremor Type and Level Analysis for PD from Videos

6.1 Introduction

To validate our research insights in Chapters 4 and 5, including the posed-based human movement disorder diagnosis framework and the movement frequency analysis, we transfer methodologies from the CP prediction project to analyze Parkinson's disease (PD), a more common and significant human movement disorder. By doing so, we also wish our research outcome can tackle different human movement disorders and improve our understanding of the underlying mechanisms across various conditions.

Parkinson's disease (PD) is the second most common progressive neurological disorder, affecting an estimated 10 million people globally [30]. It is characterized by the loss of dopaminergic neurons within the substantia nigra region of the brain, resulting in motor dysfunction [31]. Existing PD diagnosis is mainly based on the clinical assessment of PD symptoms, medical history, l-dopa and dopamine responses [32]. The clinical diagnostic accuracy is approximately 73%-84% [33], and may be affected by medical experts' subjective opinions and experiences. An automatic, efficient and interpretable PD assessment system would support clinicians

in making more robust diagnostic decisions.

Recent research in PD diagnosis with machine learning using human-centric visual, audio and movement features has shown promising results. Models based on neuroimaging [34] and cerebrospinal fluid biomarkers [35] provide an accurate diagnosis but are costly and intrusive, making them unsuitable for large-scale pre-diagnosis. Non-intrusive methods with speech [36] are limited by their generalizability due to the significant difference in language and pronunciation for patients from different geographical areas. Although gait disturbance is not typically the primary symptom of early-onset PD [37, 38], over 70% of these patients exhibit at least one form of tremor [38]. Hence, identifying Parkinson's Tremor (PT) is seen as a more generalizable approach for assisting in early PD diagnosis. To date, hand tremors-based studies mostly rely on wearable sensor data [39]. However, the use and set-up of wearable technology may be time and resource-consuming [39]. Video-based analysis with consumer-grade cameras is preferable as a more cost-effective solution without disrupting the natural behavior of the participants.

Inspired by the frequency domain modeling and pose-based human modeling discussed in Chapters 4 and 5, in this chapter, we first evaluated the performance of FAIGCN and SAFA-GCN in the PT analysis task. This was achieved by directly applying them with the minimal necessary modifications, such as changing the output layer for multiclass classification and adjusting the number of GCN layers along with the corresponding input and output channel sizes. However, the results in Table 6.1 and 6.2 show that our previous frequency domain modeling methods do not perform well in PT analysis. We hypothesize that this inadequacy is due to the frequency of hand tremor being much higher than that of infants' movements, which prevents the models from effectively distinguishing between high-frequency noise and high-frequency tremor. Additionally, FAIGCN and SAFA-GCN may lack the capacity to accurately model the more complex tremor frequency patterns.

To this end, we shift our focus to analyze PT by modeling human movement solely in the spatial domain. In this chapter, we propose a novel video-based deep

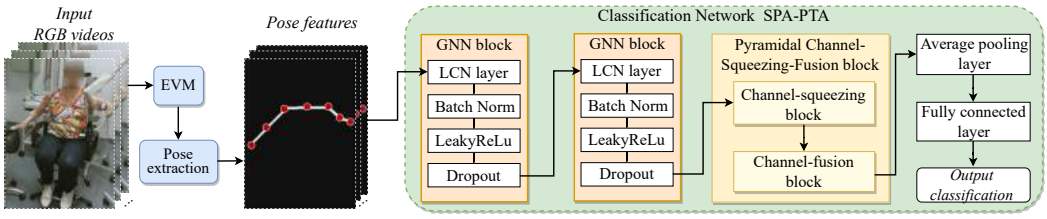


Figure 6.1: The framework of our system: we use EVM to enhance the subtle tremors in the original videos, then pass videos to the pose extraction process. We classify the extracted pose features by SPA-PTA with a novel PCSF design.

learning system, the Spatial Pyramidal Attention network for PT type and level Analysis (SPA-PTA), for PT classification and tremor severity estimation to assist the pre-diagnosis of PD with PT symptoms. Unlike the predominant systems that depend on wearable sensors [39], SPA-PTA utilizes consumer-grade cameras to record human movements. This non-intrusive method aims to offer a low-cost solution for PT classification, serving as an early warning sign for Parkinson's Disease (PD) in undiagnosed individuals. For the first time, we propose to use a novel attention module with a lightweight pyramidal channel-squeezing-fusion architecture to extract relevant PT information and filter the noise efficiently. This design aids in improving both classification performance and system interpretability. Experimental results show that our system achieves 91.3% accuracy and 80.0% F1-score in PT classification, 76.4% accuracy, and 76.7% F1-score in tremor rating classification.

6.2 Method

Fig. 6.1 shows the overview of our system. Its input is a set of videos showcasing a patient sitting in an upright posture, performing various actions such as keeping arms parallel to the ground. The human joint position features are extracted from the videos using AlphaPose [148], a state-of-the-art pose-estimation algorithm. These features are then fed into the Spatial Pyramidal Attention network for PT type and level Analysis (SPA-PTA).

6.2.1 Eulerian Video Magnification

We employ Eulerian Video Magnification (EVM) as a signal processing method [164] to enhance the subtle tremors and reduce noise and artifacts in the videos. This is motivated by previous research finding [127] that deep neural network models paid more attention to human wrists during PT classification, indicating that magnifying subtle hand and wrist motions can be beneficial for tremor feature learning. Before applying EVM, we checked the Nyquist limits [165] to examine whether our video frequency is valid for tremor analysis. Specifically, the video frame rate should be at least twice the highest frequency of tremor motions. As existing research [166] has shown that PT typically occurs between 3 and $7Hz$, our video with $30Hz$ fulfills the requirement.

6.2.2 Pose Extraction

We extract the 2D pose features from the EVM-processed videos by AlphaPose [148] since it is superior to OpenPose [1] as it demonstrates 25% improved pose estimation performance on average precision and average recall metrics in multiple datasets. We prefer 2D poses to 3D ones, as current 3D pose estimation techniques are less mature and they generally introduce noise particularity in the depth dimension [167], making them less suitable for sensitive features like tremors. We use AlphaPose to estimate 17 COCO-format [148] body keypoints and extract (x, y, c) features, where (x, y) represent the 2D coordinate and c is a confidence score that reflects the estimation accuracy. Different from using all infant's keypoints as in Chapters 4 and 5, we only utilize the top half of the body keypoints (shown in Fig. 6.5) for PT analysis. Because this approach disregards less relevant lower-body features to enhance model efficiency and reduce potential bias based on the clinical observation that PT generally occurs on the upper body, specifically on the hands and arms [168]. In addition, we omit the head joints as the participants' faces are generally obscured in medical videos to preserve their privacy.

Furthermore, we normalize the pose to mitigate bias resulting from inherent video differences. In order to mitigate global translations in the pose, we align the mean location of the neck and two hip joints as the global origin. Subsequently, all joint positions are expressed as relative values to this established origin.

6.2.3 Classification Network

We propose SPA-PTA for PT analysis by solving the PT classification task and the tremor severity estimation task. SPA-PTA is composed of two GNN blocks with a spatial attention mechanism, along with a novel pyramidal channel-squeezing-fusion block designed to learn the joint-wise relevancy.

GNN Block with Spatial Attention Mechanism: Similar to Chapter 5, we use using GNNs for PT analysis, which are effective in modeling relational data, unlike images that are in a grid structure. We model human poses as a relational graph structure $G = (V, E)$ [125], with the nodes representing the joints, and the edges representing the skeletal structure across time. Formally, $\{V = v_{p,q}\}$ represents the set of joints positions, where $v_{p,q}$ is the p-th joint at q-th frame. The set of edges, E, consists of (i) spatial edges connecting different joints in space, and (ii) temporal edges connecting the same joint across consecutive frames.

Different from using GCN [121] as in Chapters 4 and 5, we adopt the locally connected network (LCN) [124] to learn joint i's attention weight from its relationship between other joints. This method provides a way to mitigate the vanilla GCN's [121] representation capacity limitation that different joints share the same weight set. Specifically, it enables the system to learn joint i's attention from its relationship between other joints. The basic formulation is as follows:

$$\mathbf{h_i} = \sigma\left(\sum_{j \in \mathcal{N}^i} \mathbf{W}_j^i \mathbf{x}_j \hat{a}_{ij}\right), \tag{6.1}$$

where σ is an activation function, $\mathbf{W}_j^i$ is the learnable attention weight between the target node i and the related node j, $\hat{a}_{ij}$ is the corresponding element in the

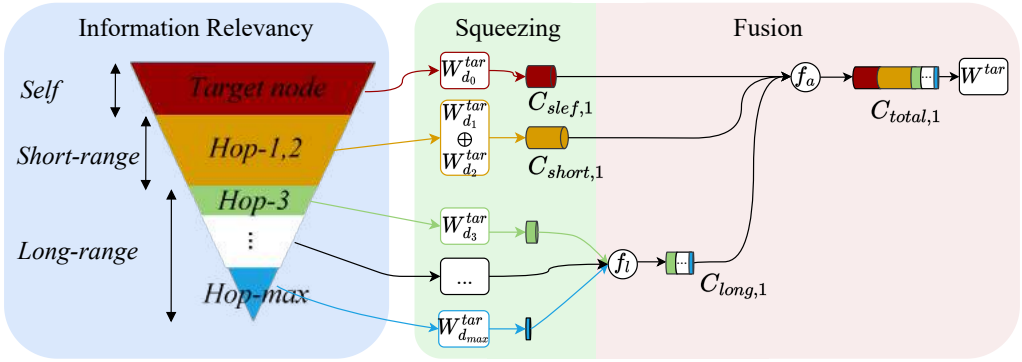


Figure 6.2: The proposed Pyramidal Channel-Squeezing-Fusion architectures.

adjacency matrix, $\mathbf{x}_j$ is the input features of node j, $\mathcal{N}^i$ is the set of connected nodes for node i, and $\mathbf{h}_i$ is the updated features of node i.

Pyramidal Channel-Squeezing-Fusion Block (PCSF): As an extension of the spatial attention module, we propose a novel inverted pyramid architecture, Pyramidal Channel-Squeezing-Fusion Block (PCSF), consisting of a *channel-squeezing block* and a *channel-fusion block* to extract relevant PT information and filter noise. This is motivated by two findings: (i) Information Gain analysis [169] shows that the information gain decreases exponentially with increasing distance between graph nodes; (ii) clinical observation [170] shows that PT usually occurs only on one side of the PD patient's upper body, such that the information relevancy between two arms should be reduced. Our proposed design does not require learnable parameters, such that it prevents overfitting problems. As shown in Fig. 6.2, the output target node i's attention weight W^i is obtained from the joint-wise weights $\{W^i_{d_0}, ..., W^i_{d_{max}}\}$ after the squeezing-and-fusion process, where d_n is the shortest distance between the target node i and the relevant node n, namely *Hop-n*. The visualization of information relevancy in Fig.6.2 guides the squeezing ratio, such that our method overcomes the limitation of the GCN [121] that each joint shares the same weight.

The Channel-Squeezing Block: We propose following squeezing operations to enhance the learning of PT-specific relevant information while filtering noise, based on our hypothesis motivated by [171, 172]. We distinguish nodes in different graph-distance by defining *hop-0* node to be the self node, *Hop-1,2* nodes to be the short-range nodes and *Hop-3*,...,*Hop-max* to be the long-range nodes. Suppose the node i is the target node, and the node j is the relevant node of i, then node j's output channel size is formulated by Eq.6.2:

$$C_{out,j} = \begin{cases} C_{in}, & |j-i| = 0 \quad , \\ pC_{in}, & 0 < |j-i| \leq 2 \quad , \\ q^{|r-i|}C_{in}, & |j-i| > 2. \end{cases} \tag{6.2}$$

where p, q are channel-squeezing ratios for *Hop-1,2* nodes and *Hop-3,..., max* nodes, respectively. $p, q \in [0, 1]$ and $p \gg q$. $C_{out,j}$ is the output channel size of node j. $|\cdot|$ denotes the graph distance between nodes.

The Channel-Fusion Block: To hierarchically combine the different range information of the target node i, we fuse the long-range features by f_l, and fuse all features by f_a:

$$\mathbf{h_i} = f_a[\mathbf{h_{self}}, \mathbf{h_{short}}, f_l(\mathbf{h_{long,k}})]\mathbf{W^i} \tag{6.3}$$

where $\mathbf{h_{long,k}}$ is the feature of the long-range node k, $\mathbf{h_{short}}$ and $\mathbf{h_{self}}$ are features of short-range nodes and self-node, respectively. $\mathbf{W^i}$ is the final weight matrix of target node i.

Implementation Details: As depicted in Fig. 6.1, our network employs two GNN blocks with output channel sizes of 64 and 128, respectively. Each block contains an LCN layer (Locally Connected Network [124]), a batch normalization layer, a LeakyReLU layer with an alpha of 0.2, and a dropout layer with 0.2 rates. Following the two GNN blocks, we employ a PCSF block, a global average pooling layer, and a fully connected layer. We adapt Cross-Entropy loss in binary

classification. To address the class imbalance in multiclass classification, we use the focal-loss [71] instead. Our optimizer of choice is Adam. The best performance of the PT binary classification task is achieved by a learning rate of 0.01 (decays by 0.1), a batch size of 8, and a dropout of 0.2, at 500 epochs.

6.3 Dataset

We test our system using the TIM-TREMOR dataset [173], which is an open dataset consisting of 910 videos of 55 individuals performing 21 tasks and videos range from 18 seconds to 112 seconds. The RGB video resolution is 1920 $\times$ 1080 with a sampling rate of 30 frames per second. Patients were recruited from the outpatient clinic of the Department of Clinical Neurophysiology of the Leiden University Medical Center, Netherlands. Diagnostic labels about different tremors were attributed by a neurologist in movement disorders on the basis of available medical information. Tremor severity was evaluated by a experienced clinician using the Bain and Findley Tremor Clinical Rating Scale [174]. Specially, the tremor severity of each arm was scored on a scale ranging from 0 to 7 (0: no tremor, 1-3: mild tremor, 4-6: moderate tremor, 7: severe tremor).

There are 572 videos depicting various forms of tremors, including 105 for Parkinsonian Tremor (PT), 182 for Essential Tremor (ET), 88 for Functional Tremor (FT), and 197 for Dystonic Tremor (DT). An additional 60 videos (NT) were recorded without convincing tremors during the assessment. The test 278 videos have inconclusive tremor classification results and have been labeled as 'Other.' For the tremor rating labels, eight levels from level 0 to 7 are assigned to the individual's left and right hands. To ensure that there is only one label per video and preserve the characteristics of the video, we combine the labels for individual left and right hands by taking the maximum value of both hands.

Ethical approval: Approval of the TIM-TREMOR dataset was obtained from the University Leiden University Medical Center ethics committee. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

6.4 Experiments

To assess the efficacy of our proposed method, we conducted validation testing on two separate evaluation exams: the PT classification exam and the tremor rating estimation exam. We carried out our experiments using a Ubuntu 18.04 PC with an NVIDIA 3080. The GPU memory usage for training was minimal, averaging just 1.46 gigabytes. The training process for the TIM-TREMOR dataset took approximately ten hours for the PT classification task, and twelve hours for the tremor rating estimation exam. They include the processes of EVM and extraction of human pose features from RGB videos. In terms of real-time application, the PT classification or tremor rating estimation of a 33 seconds video with 1000 frames only took around 48 seconds each, which is a feasible time for interactive diagnosis.

Setup: We eliminate inconsistent videos to minimize data noise, specifically, videos that only capture motion tasks for a limited number of participants. For the tremor-type classification task only, we remove the videos with uncertain tremor-type labels of 'other'. Next, we follow the previous PD analysis work [175] to clipp each video into 100-frame samples, and the number of samples is determined by the duration of the consecutive frames in which the participant was visible and not obscured. Each sample was assigned the label of the original video and treated as an individual sample. To be consistent with the previous evaluation protocol on PD analysis [56, 175], we use the same voting system rather than the clipping-and-fusion technique in Chapter 5 to obtain the video-level classification results, which increases the system's robustness and augments the training sample size. We evaluate our proposed system through individual-based leave-one-out cross-

	Binary Classification				Multiclass Classification			
Method	AC	SE	SP	F1	AC	SE	SP	F1
ST-GCN[125]	84.6	71.4	87.5	62.5	64.1	64.8	90.7	64.1
CNN-Conv1D	76.9	57.1	81.3	47.1	56.4	54.2	88.3	53.1
Decision Tree	69.2	57.1	71.9	40.0	51.3	49.4	87.6	36.7
SVM [75]	64.1	57.1	65.6	36.4	46.2	44.6	86.2	44.3
FAIGCN [26]	82.1	71.4	84.4	58.8	61.5	64.0	90.1	63.8
SAFA-GCN	84.6	71.4	87.5	62.5	66.7	67.3	91.5	67.9
Ours - full	**92.3**	**85.7**	**93.8**	**80.0**	**71.8**	**71.3**	**92.5**	**72.5**
w/o PCSF	87.2	85.7	87.5	70.6	66.7	67.6	91.4	66.7
w/o Attention	82.1	71.4	84.4	58.8	61.5	63.1	90.0	62.4
w/o Attention & EVM	79.5	71.4	81.3	55.6	59.0	59.1	89.5	58.5

Table 6.1: The comparisons on the tremor type classification task.

validation. It means each subclips for a single individual is used for testing and excluded from the training set for each iteration. The subclips for each individual are never separated by the training or testing set.

Evaluation Metrics: We report the mean values calculated among all leave-one-out cross-validations with the following metrics: accuracy (AC), sensitivity (SE), specificity (SP), and F1-Score for the binary classification; AC, macro-averaged F1-score, SE and SP for the multiclass classification.

6.4.1 Tremor Type Classification

For this experiment, we first evaluate our system on the binary classification that distinguishes PT labels from non-PT labels, and achieve 91.3% accuracy and 80.0% F1-score. In addition, we validate our method on a more complex multiclass classification task for classifying five types of tremors (PT, ET, DT, FT and NT). Our final system's per-class tremor type multiclass classification performance is shown in Fig. 6.3. It shows a fairly balanced performance on classifying PT, ET, DT and NT, while FT has a lower SE and F1-score, which may be caused by the smallest number of samples in this class. Moreover, the corresponding confusion matrices of the two tasks are displayed in Fig. 6.4.

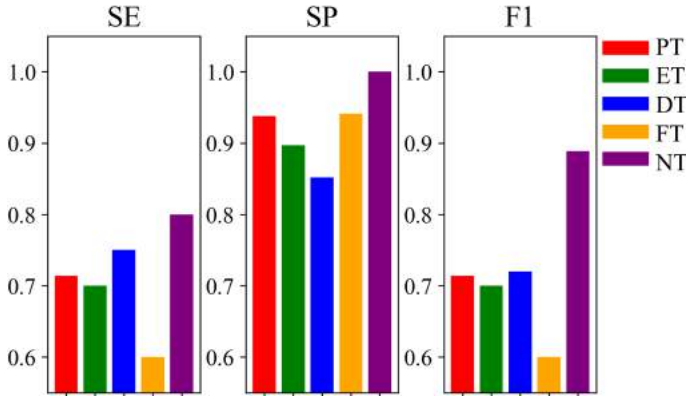


Figure 6.3: Per-class multiclass tremor type classification results.

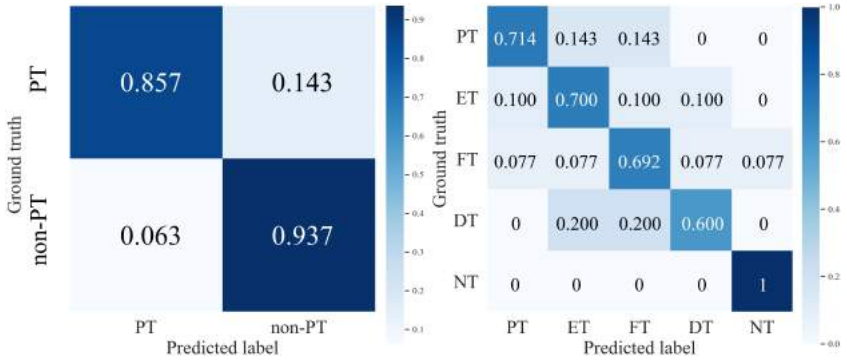


Figure 6.4: Confusion matrices for PT classifications: (Left) binary; (Right) multiclass.

Comparison with Baseline Methods: As this paper is the first work that provides the individual-level evaluation results, we implemented the following video-based PT classification baselines to evaluate the effectiveness of our system: (i) ST-GCN [125]: a spatial-temporal GCNs for human pose data classification; (ii) CNN with 1D convolutional layers (CNN-Conv1D) [75]; (iii) Decision Tree (DT); (iv) Support Vector Machine (SVM) [75]; (v) FAIGCN [26] in Chapter 4; (vi) SAFA-GCN in Chapter 5. Note that all baseline methods apply the same EVM and pose extraction design. The results of our proposed SPA-PTA and baselines are summarized in Table 6.1.

The binary classification result shows that our full system consistently outperforms all other methods in all evaluation metrics. Our AC, SE, SP, and F1 achieve over 80% on leave-one-out cross-validation, demonstrating the effective-

ness and stability of our system in this binary classification task. It is noticeable that our system performs better with only spatial convolution instead of a deeper spatial-temporal convolution design like ST-GCN [125]. In addition, our previous model, FAIGCN, which relies solely on frequency domain modeling, exhibits a performance decline compared to ST-GCN. We hypothesize that this limitation arises because the frequency of hand tremor is significantly higher than that of infants' movements, impeding the models' ability to effectively differentiate between high-frequency noise and high-frequency tremor. Furthermore, our SAFA-GCN, which combines both spatial and frequency domain modeling, demonstrates only a marginal improvement in the multiclass classification task. These outcomes suggest that our initial approaches to frequency domain modeling may not be directly applicable to modeling high-frequency movements. Moreover, our earlier models may not have the capacity to accurately represent the more intricate tremor frequency patterns. These results motivated us to propose a spatial domain model with more capacities. The outcome of our full SPA-PTA supports that the suggested PCSF block effectively enhances classification reliability and reduces the risk of overfitting in small datasets.

While the full system is initially designed for binary classification, it presents effectiveness and generalizability in the multiclass classification task, surpassing existing methods. A small difference between AC, SE, and SP shows that our system performs consistently and effectively at identifying the positive samples and excluding the negative ones. The high macro-average SP exhibited trustworthy effectiveness in correctly recognizing individuals who have a specific type of tremor without wrongly recognizing it as other types of tremor.

Ablation Studies: We conduct an ablation analysis to assess the effectiveness of the EVM, PCSF block and the entire attention module. From the lower parts of Table 6.1, the positive effect of the PCSF block and attention module can be illustrated by the decrease in metrics when either the PCSF block or the entire

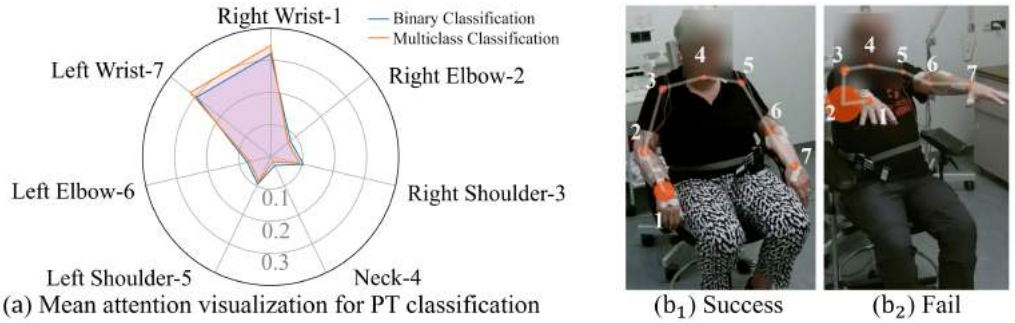


Figure 6.5: (a) The average skeleton joints attention across all cross-validations in the PT classification experiment; (b) The attention visualization at a (b_1) successfully classified frame, and (b_2) unsuccessful classified frame. The joint labels in (b) correspond to (a);

attention module is removed in the two classification tasks. Also, we find that the basic GNN architecture without attention performs better than the CNN-Con1D model for both classification tasks. It highlights the efficacy of learning human pose features in the graph domain as opposed to the Euclidean domain. Besides, the variant of 'ours without attention' performs slightly better than 'ours without attention and EVM preprocessing', indicating that the use of EVM could effectively enhance tremors.

Model Interpretation: We present the visualization for the average attention value of each body keypoint in Fig. 6.5a. It is interpreted as the importance level our system considers during the classification process. Our analysis reveals that the attention value is significantly highest on the 'Right Wrist' and 'Left Wrist', which suggests that our system prioritizes the wrists' movements during the task performance. Furthermore, the value associated with the 'Neck' is significantly lower than other keypoints. It may be explained by the fact that the participants remained seated during the video recording, resulting in a minimal global variance of the neck joint throughout the experiment.

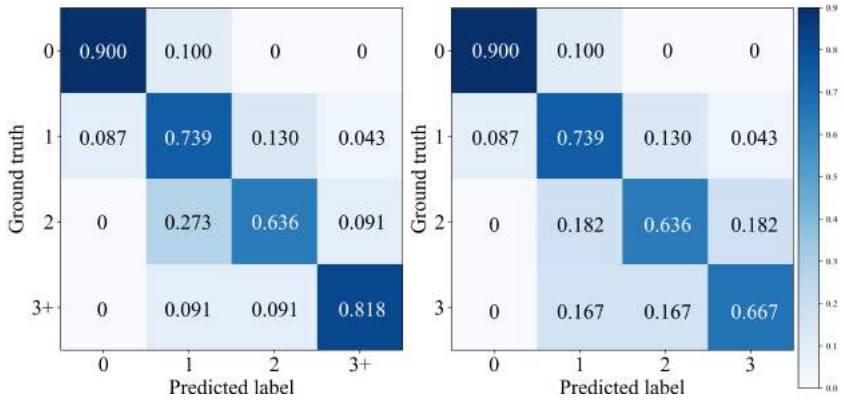


Figure 6.6: Confusion matrices for tremor rating estimation: (Left) [0,1,2,3+]; (Right) [0,1,2,3].

6.4.2 Tremor Rating Estimation

For this experiment, we train SPA-PTA with different tremor rating labels without any further implementation (e.g., converting the classification layer to a regression layer) to validate our system performance in the tremor rating estimation task. Since the data with tremor ratings 4 and above is insufficient for training via leave-one-out cross-validation (i.e., only 5 individuals out of 55), we validate our system on two different classification settings: (1) Classifying ratings [0,1,2,3] (2) Classifying ratings [0,1,2,3+]. The latter is generally a more challenging task since the imbalanced data of '3+' rating brings bias, compared to the former which does not contain such data.

Comparison with Baseline Methods: We compare our SPA-PTA to the same baselines in the tremor-type classification task as shown in Table. 6.2. SPA-PTA significantly outperforms the baselines by achieving a macro-average AC of 76.4%, SE of 77.3%, SP of 91.6%, and F1-score of 76.7%. An interesting finding is that the machine learning-based method Decision Tree achieves similar performance to two deep learning-based baselines (i.e., ST-GCN and CNN-Conv1D). It may inform us to further tackle the challenge of improving the deep learning models in a relatively small dataset. In addition, although our current model does not

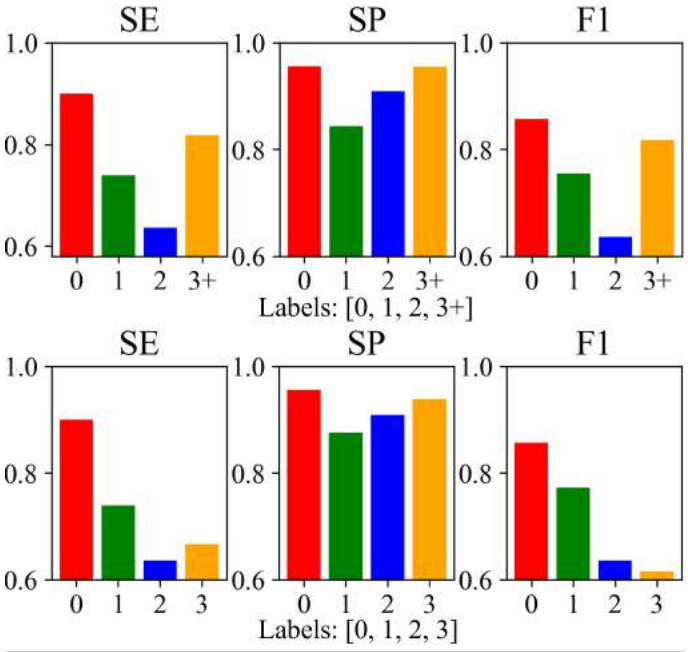


Figure 6.7: Per-class tremor rating estimation results.

show strong robustness in the tremor rating estimation task, the ablation studies from the rows of 'Ours' in Table 6.2 still demonstrate the effectiveness of our PCSF layer and the attention mechanism design. It shows the potential of improving our model and system performance with a more specific architecture design with a more extensive dataset. Moreover, it is not surprising that our CP prediction models, FAIGCN and SAFA-GCN, show similar performance rankings to the tremor type classification task. This suggests that spatial modeling with a larger model capacity might be a more effective approach for modeling high-frequency movements.

Classification labels	$[0, 1, 2, 3]$				$[0, 1, 2, 3+]$			
Method	AC	SE	SP	F1	AC	SE	SP	F1
ST-GCN[125]	67.3	68.1	89.0	66.5	68.0	67.7	90.5	65.7
CNN-Conv1D	60.0	59.8	86.5	58.7	60.0	60.5	87.9	58.3
Decision Tree	54.5	55.3	85.2	54.6	52.0	53.0	86.0	51.3
SVM [75]	49.1	41.1	81.5	43.8	48.0	49.5	85.2	47.1
FAIGCN [26]	63.6	64.5	87.4	64.7	64.0	64.1	87.0	63.0
SAFA-GCN	70.9	72.9	89.9	72.3	68.0	67.5	88.5	66.9
Ours - full	**76.4**	**77.3**	**91.6**	**76.7**	**74.0**	**73.5**	**92.0**	**72.0**
w/o PCSF	70.9	71.5	89.7	70.7	70.0	68.6	90.5	68.2
w/o Attention	65.5	65.6	88.2	64.8	66.0	65.2	89.5	63.9
w/o Attention & EVM	63.6	64.8	87.6	63.3	64.0	64.1	88.9	62.5

Table 6.2: The comparisons on the tremor rating task.

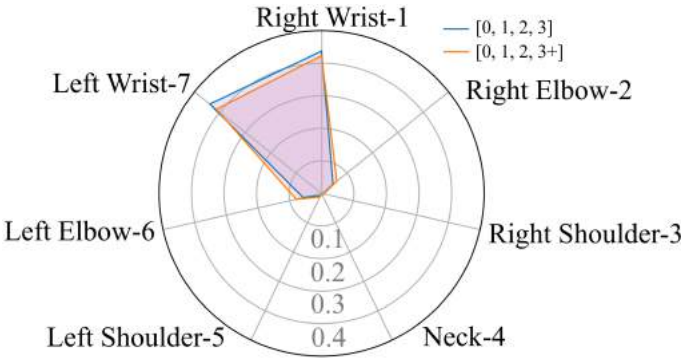


Figure 6.8: The average skeleton joint attention across all cross-validations in tremor rating estimation task

Ablation Studies: Consistent results at the bottom of Table 6.2 from the same ablation design as for the PT classification task validate the effectiveness of each system component.

Model Interpretation: We similarly visualize the average skeleton joints attention across all cross-validation sets in Fig. 6.8. Two different data preprocessing approaches provide similar attention results, while the weights obtained by grouping [0,1,2,3] slightly more contribute to 'Right Wrist' and 'Left Wrist'. This may be due to the increased proportion of low tremor rating videos in this approach compared to grouping [0,1,2,3+]. In addition, we notice that the attention weight distribution of the tremor rating estimation exam is similar to that of the PT classification exam, while the former aggregates more attention on the 'Right Wrist' and 'Left Wrist' than other joints.

6.4.3 Pose Estimation Evaluation

To evaluate the effectiveness of AlphaPose and quantify the pose estimation error, we conduct the following experiments:

Quantitative Comparison with Ground Truth Data: To quantify the pose estimation error from different methods, we employ the Lagrangian hand-tremor

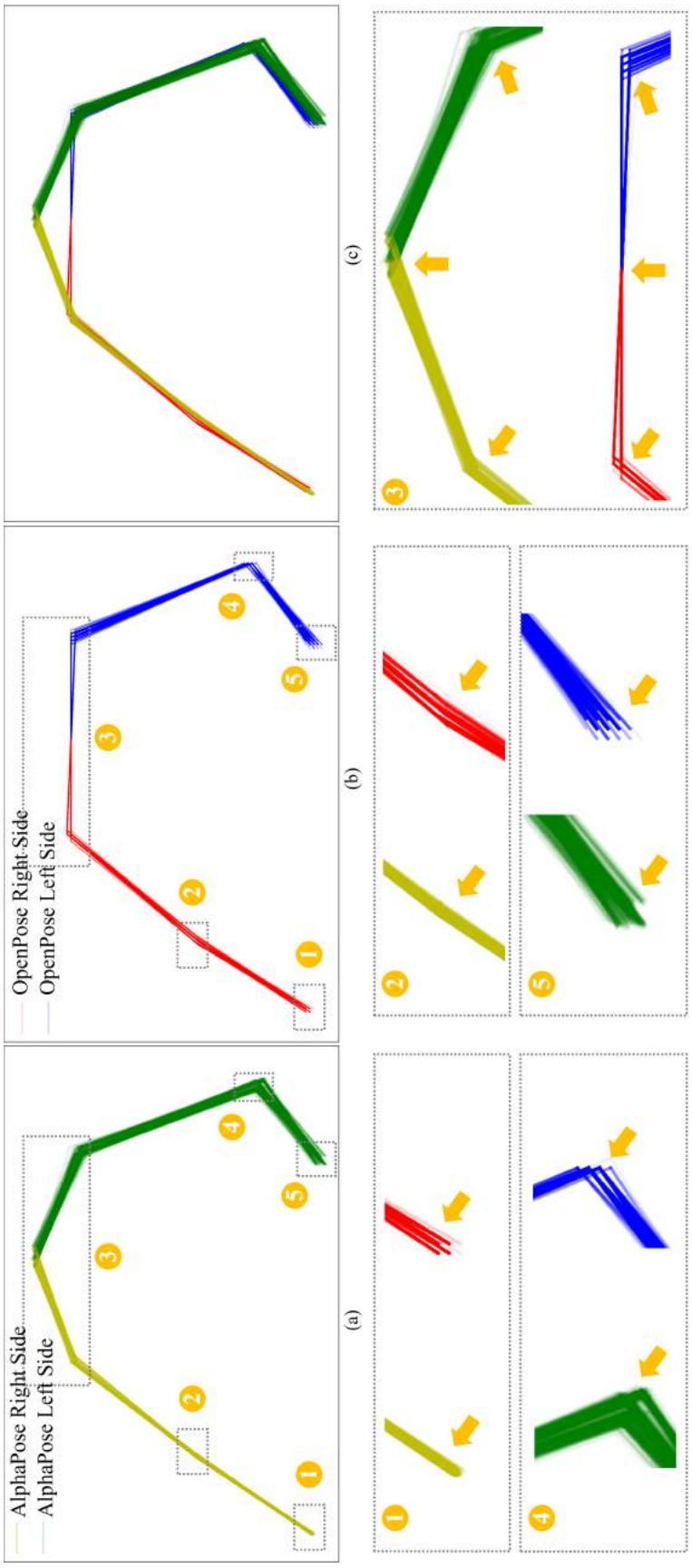
Task	AlphaPose	OpenPose
Rest	**0.812**	0.881
Rest in supination	**0.834**	0.930
2 Hz higher	**0.605**	0.635
2 Hz lower	**0.617**	0.622
Counting	**0.729**	0.790
Finger tapping	**0.576**	0.687
Playing piano	**0.752**	0.906
Months backward	**0.814**	0.838
Top top	**0.786**	0.823
Thumbs up	**0.844**	0.960
Average MAE	**0.737**	0.807

Table 6.3: MAE comparison between AlphaPose features and OpenPose on the top-10 best-performing tasks. Better performance with lower MAE is in bold.

frequency estimation method [173] to compare MAE (Mean Absolute Error) of the hand tremor frequencies estimated by AlphaPose and conventional OpenPose features [127] with Ground Truth (GT) frequency obtained from accelerometer data. As suggested in [173], tremor frequency calculated from reliable estimated pose features should be close to (i.e., ideally within 1 HZ difference) the GT accelerometer data frequency. The MAE from Table 6.3 indicates that AlphaPose consistently outperforms OpenPose on all listed tasks.

Qualitative Pose Visualization and Comparison: The visualizations in Fig. 6.9 and the reference video images in Figure 6.10 show that AlphaPose outperforms OpenPose in estimating joint positions. This is supported by the smoother trajectory lines of AlphaPose, which are depicted by the transparent colored lines. Sub-figures 1 to 5 in Fig. 6.9 demonstrate AlphaPose's ability to track joint movement fluidly. Specifically, in sub-figure 5, AlphaPose demonstrates a hand trajectory that aligns more closely with the anticipated tremor pattern, which contrasts with OpenPose's intermittent jumping trajectory. This consistency suggests that AlphaPose may be more reliable for tasks related to PT classification. Furthermore, on the patient's right side, particularly in sub-figures 1 and 2, AlphaPose yields more consistent and stable outcomes, reflecting the patient's condition of

resting with observable tremors only in the left hand, as corroborated by Figure 6.10. Finally, the neck joint position of OpenPose is estimated by the mean point of both shoulders, which is less accurate than the estimated neck joint position of AlphaPose [148].



.

Figure 6.9: The estimated pose comparison between AlphaPose and OpenPose for a sitting and resting PD patient with clinically identified PT on the left side of the body. (a), (b) and (c) are the estimated poses of an example video from AlphaPose, OpenPose, and both, respectively. Each colored line with 0.05 transparency represents the connection between joints estimated in each frame. Numbers 1 to 5 correspond to specific joints' local scaling for intuitive comparison. The raw video frames are referenced in Fig. 6.10

Figure 6.10: The raw videos referenced in Fig. 6.9 consist of consecutive images captured at intervals of 5 frames, approximately every 0.167 seconds. The lower-right image is an aggregation of five transparent hand images, where the green dot shows the estimated trajectory of the left wrist joint during tremor.

Method	AC	SE	SP	F1	AC	SE	SP	F1
	Tremor-type Binary				Tremor-type Multiclass			
AlphaPose	**92.3**	**85.7**	**93.8**	**80.0**	**71.8**	**71.3**	**92.5**	**72.5**
OpenPose	**92.3**	**85.7**	**93.8**	**80.0**	69.2	69.3	92.0	70.2
	Tremor-level [0,1,2,3+]				Tremor-level [0,1,2,3]			
AlphaPose	**76.4**	**77.3**	**91.6**	**76.7**	**74.0**	**73.5**	**92.0**	**72.0**
OpenPose	72.7	74.0	90.4	73.6	72.0	72.5	90.4	70.1

Table 6.4: The comparisons on the influence of classification performance between AlphaPose and OpenPose.

Classification Performance Comparison: We compare the effectiveness of AlphaPose and OpenPose by evaluating their impacts on the system classification performance. Table 6.4 demonstrates that using AlphaPose features results in a remarkable and consistent improvement over OpenPose features of approximately $1-3\%$ across the classification tasks except for the binary tremor-type classification. These results highlight the precision of AlphaPose in delivering better pose-based features for classification tasks.

In this study, we utilize the pre-trained AlphaPose model, opting not to retrain it due to the absence of GT 2D pose position annotations within our dataset. The robust generalization capability of the pre-trained AlphaPose model, as evidenced by its superior performance across multiple diverse and complex benchmark datasets [148], affirms its suitability for our task. In the future, we are interested in comparing the performance between pretrained and tremor-specific pose estimation models. This will entail the collection of the necessary GT data to train a model adept at detecting the subtle nuances characteristic of tremor movement patterns.

6.5 Generalizing to the CP Prediction Task

To further evaluate the generalization ability of SPA-PTA, we directly apply it to the CP prediction task with the minimal necessary modifications, such as utilizing all body keypoints rather than only upeer body keypoints, and optimizing the number of GNN layers along with the corresponding input and output channel

	Method	AC	SE	SP	F1
LOOCV	SPA-PTA	97.37	83.33	100.00	90.91
	SAFA-GCN	97.37	83.33	100.00	90.91
	FAIGCN [26]	97.37	83.33	100.00	90.91
5-fold CV	SPA-PTA	92.11	83.33	93.75	76.92
	SAFA-GCN	97.37	83.33	100.00	90.91
	FAIGCN [26]	94.74	83.33	96.88	83.33
3-fold CV	SPA-PTA	92.11	83.33	93.75	76.92
	SAFA-GCN	94.74	83.33	96.88	83.33
	FAIGCN [26]	92.11	83.33	93.75	76.92

Table 6.5: Comparing SPA-PTA with our previous CP prediction models on the RVI-38

sizes, and validate the optimized SPA-PTA on the RVI-38 dataset. As shown in Table 6.5, it is not surprising that our SPA-PTA obtains the lowest performance among all three models. This is because the PCSF module, which is specifically designed for PT analysis, does not theoretically align with CP movement analysis. Specifically, the left side of infant's body should be highly relevant to the right side [152], but the PCSF module reduces such relevancy as PT typically occurs only on one side of a patient's upper body [170]. Therefore, this suggests that the design of biomedical engineering models should fully incorporate clinical observations and guidance, rather than focusing solely on improving model capacity.

6.6 Limitations and Discussions

Our findings about PT analysis are preliminary, and the limited number of people with PT and the limited range of tremor levels included in this work may affect the generalizability of the results. One of our future directions is to evaluate our models using data collected from a larger and more diverse group of PD patients, covering a more balanced tremor-type distribution and a wider range of tremor severity ratings. Up-scaling the study is crucial for developing more robust models and for enhancing the overall applicability and validity of the framework we have presented. In addition, annotating the dataset based on PT severity estimation performance by different scales, such as the MDS-UPDRS3, by experienced

raters will enable us to improve the robustness of our model via more fine-grained designs in the future. Besides, our current system performance is still challenged by pose estimation algorithm error, such as depicted in Fig. 6.5b. The attention of our system is incorrectly influenced by the inaccurate position detection of the right elbow and blurred right shoulder joints. Moreover, our SPA-PTA does not employ the clipping-and-fusion technique described in Chapter 5 to interpret the temporal characteristics of PT. In addition to following the existing PD analysis protocol [175], this decision is prompted by the concern that the overlapping frame length might unevenly split some tremor cycles between subclips, potentially compromising the accuracy of subsequent analyzes. In future studies, we plan to explore alternative methods for capturing and analyzing the temporal dynamics of PT without relying on overlapping frames. This may involve developing new algorithms that can more accurately classify and quantify tremors throughout the entire video.

6.7 Conclusion

Our method effectively identifies PT in PD patients from consumer-grade videos. The validity of our proposed system on both PT classification and tremor severity estimation tasks demonstrates that our method is extensible in PT-related analysis. Our non-intrusive system only relies on consumer-grade videos as input, so it offers a potential cost-effective solution for supporting the pre-diagnosis of PD in regions with inadequate medical resources. This work also could be used for remote PD supplementary assessment in special situations to reduce the stress of the healthcare system (e.g., COVID-19 pandemic). Moreover, our system demonstrates the potential to automatically monitor PT symptoms during daily activities to support PD pre-diagnosis.

Statements and Declarations

H. Shum received support from the EPSRC NortHFutures project (Ref: EP/X031012/1). S. D. Din has received support from Innovative Medicines Initiative 2 Joint Undertaking (Ref: 820820 Mobilise-D, 853981 IDEA-FAST), NIHR Newcastle, Newcastle upon Tyne Hospitals NHS Foundation Trust, Cumbria Northumberland and Tyne and Wear NHS Foundation Trust.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we applied and developed state-of-the-art GNNs with additional clinical guidance (e.g., HCSF, masked frequency-attention) for developing automated disease diagnosis systems that are accurate, robust, and whose decisions can be explained by humans. We proposed several attention mechanisms to improve system interpretability and robustness. We demonstrated such automated diagnostic tools can not only be used to assist clinicians in making more comprehensive and precise diagnoses, but also have the potential to offer low-cost diagnostic support for regions with limited clinical resources.

For CP prediction, our final system integrates the spatial attention, frequency attention, and a clipping-and-fusion method to strengthen the prediction reliability and interpretability. Our system visualizes important human joints, frequency bands and time ranges in CP prediction to support clinicians in making accurate and robust decisions. We also supplement the MINI-RGBD dataset and RVI-38 dataset with more accurate posture features and provide a performance benchmark analysis of leading methods.

For PD analysis, our method effectively identifies PT in PD patients from consumer-grade videos. The validity of our proposed system on both PT classifica-

tion and tremor severity estimation tasks demonstrates that our method is extensible in PT-related analysis. Our non-intrusive system only relies on consumer-grade videos as input, so it offers a potentially cost-effective solution for supporting the pre-diagnosis of PD in regions with inadequate medical resources. This work could also be used for remote PD supplementary assessment in special situations to reduce the stress of the healthcare system (e.g., the COVID-19 pandemic). Moreover, our system demonstrates the potential to automatically monitor PT symptoms during daily activities to support PD pre-diagnosis.

7.2 Limitations and Future Directions

Firstly, the most significant limitation of our work on the dataset size of human movement disease videos, which is also an important future direction. The conclusions drawn from our current analysis of CP prediction and PT are tentative due to the narrow spectrum of tremor intensities examined, and the small sample sizes of infants diagnosed with CP and individuals experiencing PT. These factors may limit the generalizability of our findings. To overcome this limitation, we highly recommend that researchers interested in this field participate in the data collection and annotating process, which could significantly advance the development of the research community. For example, annotating the dataset based on PT severity estimation performance by different scales, such as the MDS-UPDRS3, by experienced raters will enable researchers to improve the robustness of model in the future via more fine-grained designs. In addition, an important future direction involves augmenting the dataset with contributions from a wider range of clinical practitioners, as well as developing effective and robust data augmentation tools that are tailor-made for analyzing the specific human movement disorders.

Additionally, enhancing the interpretability of models for human movement diseases analysis (e.g., CP prediction and PD analysis) is a critical aspect of future research. Our experiments with spatial-attention and frequency-attention designs

demonstrate the effectiveness of the attention mechanism in improving both the performance and interpretability of DNNs, which are particularly important for diagnosing human movement diseases. Moreover, while we encountered challenges in developing self-attention-based systems [92, 101] due to the limited size of our dataset, we recommend that future research with adequate data explore clinical-guided self-attention models, which offer greater capacity and flexibility in designing attention maps. Besides, the exploration of more advanced mechanisms is encouraged as they can replace or be used in conjunction with traditional attention mechanisms. For example, the newly proposed RetNet [176] claims that the model can achieve parallel training, low inference cost and strong performance. In addition, the concept of 'chain of thought' prompting [177, 178], where models exhibit their reasoning process transparently and improve problem-solving in a zero-shot context, exemplifies the symbiotic relationship between interpretability and performance. This may implicitly reveal the effectiveness of the attention mechanism, but it is essential to build a solid and complete theoretical framework to support these current explorations of attention mechanisms in various fields. A rigorous theoretical foundation can enhance confidence in the universality of attention mechanisms and promote their wider adoption in practical applications.

Besides, our research experience highlights the effectiveness of frequency domain modeling in analyzing infant movement. However, this frequency analysis method does not yield ideal results in PT analysis tasks. We hypothesize that this limitation stems from the significantly higher frequency of hand tremors compared to infants' movements, which hampers the models' ability to effectively distinguish between high-frequency noise and high-frequency tremor. This hypothesis is subjective and preliminary, future research exploring modeling tremors in the frequency domain or in both spatial-temporal and frequency domains by considering the time-frequency consistency [25] could be a very interesting and significant research direction for understanding the spatial and frequency characteristics of different human movement disorders.

Moreover, our experiment results in both CP prediction and PT analysis tasks demonstrate the superiority of GNNs over CNNs in modeling non-Euclidean structured data, such as human pose. This is supported by the graph-based nature of GNNs, which allows for a more flexible and accurate representation of complex relationships inherent in human poses. By designing graph typologies in the spatial, temporal, and frequency domains, GNNs can facilitate a deeper understanding of the spatial relationships between different body parts, the temporal relationships between different frames, and the frequency domain relationships between different frequency bands, thus enhancing the accuracy of the classification system. Moving forward, an interesting GNN research direction involves exploring the optimal design about integrating GNNs with MLPs, and/or State Space Models (e.g., Mamba [179]), and/or self-attention mechanisms [101]. This is because, theoretically, self-attention is a special case of GNN where every node is connected with others [180]. In addition, designing GNNs to be more adaptable for other deep learning techniques, such as Mamba [179]—a powerful challenger to self-attention—is expected to be a highly meaningful research topic in the future. Besides, we expect that investigating the frequency domain GNNs with capabilities to reduce the computational cost of frequency transformation (e.g., FFT) would also be a valuable research direction. Because such designs could accommodate adequate expressiveness and achieves much lower complexity, effectively and efficiently accomplishing classification and regression tasks. The corresponding theoretical proof can be found in [181].

Furthermore, given that our human movement disease analysis framework relies on the quality of pose estimation algorithms, such as OpenPose and AlphaPose, which are trained using data from healthy adults, we aim to enhance this framework through domain adaptation techniques [182] or the generation of synthetic data to further improve pose quality. This is crucial, as the pose distributions of patients with movement disorders inevitably differ from those of healthy individuals. Furthermore, we plan to upgrade the 2D pose estimation to a more accurate

and robust 3D pose estimation method, which will be particularly beneficial for our research, especially for the CP prediction task. The depth information provided by 3D pose estimation can enhance the model's capacity and pose interpretability, making it more robust to occlusion issues.

Lastly, future research should involve clinical professionals more deeply in the iterative refinement process of these models. By developing enhanced visualization tools that integrate clinician feedback, we aim not only to make model predictions more transparent but also to leverage expert insights for identifying additional spatiotemporal features. Enhanced feature extraction, achieved through methodologies like active learning or weakly supervised learning, would allow for a more detailed examination of the temporal dynamics present in physical assessments. Such advancements are expected to improve the precision with which clinically meaningful features are identified and applied within predictive frameworks.

Bibliography

[1] 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.

[2] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Muller-Felber, and A. S. Schroeder, "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.

[3] R. L. Sacco, "Neurology: challenges, opportunities, and the way forward," *Neurology*, vol. 93, no. 21, pp. 911–918, 2019.

[4] A. Nitkunan, J. Lawrence, and M. Reilly, "Association of british neurologists: Uk neurology workforce survey," *Advances in Clinical Neurosicence and Rehabilitaiton*, 2020.

[5] A. Mahajan and L. C. Shih, "Introduction to diagnostic challenges in movement disorders," *Seminars in Neurology*, vol. 43, no. 01, pp. 002–003, 2023.

[6] Z. Sun, J. Liu, Q. Ke, H. Rahmani, M. Bennamoun, and G. Wang, "Human action recognition from various data modalities: A review," *CoRR*, vol. abs/2012.11866, 2020.

[7] Z. Guo and H. Wang, "A deep graph neural network-based mechanism for social recommendations," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2776–2783, 2020.

[8] Y. Salehi and D. Giannacopoulos, "Physgnn: A physics–driven graph neural network based model for predicting soft tissue deformation in image–guided neurosurgery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 282–37 296, 2022.

[9] J. J. Q. Yu and J. Gu, "Real-time traffic speed estimation with graph convolutional generative autoencoder," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3940–3951, 2019.

[10] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: a survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[11] D. R. Patel and et al., "Cerebral palsy in children: a clinical overview," *Transl Pediatr*, vol. 9, pp. S125–S135, 2020.

[12] I. Novak, C. Morgan, L. Adde, and et al., "Early, accurate diagnosis and early intervention in cerebral palsy: Advances in diagnosis and treatment," *JAMA Pediatr*, vol. 9, no. 171, pp. 897–907, 2017.

[13] V. de Graaf-Peters and et al., "Development of postural control in typically developing children and children with cerebral palsy: Possibilities for intervention?" *Neuroscience and Biobehavioral Reviews*, vol. 31, no. 8, pp. 1191–1200, 2007.

[14] T. A. Hope and et al., "Selecting and assessing quantitative early ultrasound texture measures for their association with cerebral palsy," *IEEE Transactions on Medical Imaging*, vol. 27, no. 2, pp. 228–236, 2008.

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

[16] 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.

[17] Y. Gao, Y. Long, Y. Guan, A. Basu, J. Baggaley, and T. Ploetz, "Towards reliable, automated general movement assessment for perinatal stroke screening in infants using wearable accelerometers," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, 2019.

[18] K. Raghuram and et al., "Automated movement analysis to predict motor impairment in preterm infants: a retrospective study," *J Perinatol.*, vol. 39, pp. 1362–1369, 2019.

[19] A. S. Schroeder, N. Hesse, R. Weinberger, and et al., "General movement assessment from videos of computed 3d infant body models is equally effective compared to conventional rgb video rating," *Early Human Development*, vol. 144, p. 104967, 2020.

[20] 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.

[21] 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.

[22] 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.

[23] R. Morais and et al., "Robust and interpretable general movement assessment using fidgety movement detection," *IEEE J. Biomed. Health Inform.*, pp. 1–12, 2023.

[24] 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*, vol. 24, no. 11, pp. 1225–1234, 2016.

[25] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3988–4003, 2022.

[26] H. Zhang, H. P. H. Shum, and E. S. L. Ho, "Cerebral palsy prediction with frequency attention informed graph convolutional networks," in *EMBC*, 2022.

[27] 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 *Computer Vision - ECCV 2018 Workshops*. Springer International Publishing, 2018.

[28] 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.

[29] 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.

[30] P. Chopade and et al., "Alzheimer's and parkinson's disease therapies in the clinic," *Bioeng Transl Med.*, 2023.

[31] T. R. Mhyre and et al., "Parkinson's disease." *Sub-cellular biochemistry*, vol. 65, pp. 389–455, 2012.

[32] S. A. Mostafa and et al., "Examining multiple feature evaluation and classification methods for improving the diagnosis of parkinson's disease," *Cognitive Systems Research*, vol. 54, pp. 90–99, 2019.

[33] G. Rizzo and et al., "Accuracy of clinical diagnosis of parkinson disease: a systematic review and meta-analysis," *Neurology*, vol. 86, no. 6, pp. 566–576, 2016.

[34] L. Zhang and et al., "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers in neuroscience*, vol. 14, p. 779, 2020.

[35] W. Wang and et al., "Early detection of parkinson's disease using deep learning and machine learning," *IEEE Access*, vol. 8, pp. 147 635–147 646, 2020.

[36] J. C. Vásquez-Correa and et al., "Multimodal assessment of parkinson's disease: a deep learning approach," *JBHI*, vol. 23, no. 4, pp. 1618–1630, 2018.

[37] J. M. Hausdorff, "Gait dynamics in parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling," *Chaos*, vol. 19, no. 2, p. 026113, 2009.

[38] P. Rizek and et al., "An update on the diagnosis and treatment of parkinson disease," *CMAJ*, vol. 188, no. 16, pp. 1157–1165, 2016.

[39] M. D. Hssayeni and et al., "Wearable sensors for estimation of parkinsonian tremor severity during free body movements," *Sensors*, vol. 19, no. 19, p. 4215, 2019.

[40] 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," *Developmental medicine and child neurology*, vol. 47, no. 8, pp. 571–576, 2005.

[41] H. F. R. 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, p. 1361–1363, 1997.

[42] E. Ricci, C. Einspieler, and A. K. Craig, "Feasibility of using the general movements assessment of infants in the united states," *Phys Occup Ther Pediatr.*, vol. 38, no. 3, pp. 269–279, 2018.

[43] A. K. L. Kwong and et al., "Predictive validity of spontaneous early infant movement for later cerebral palsy: a systematic review," *Developmental Medicine and Child Neurology*, vol. 60, no. 5, pp. 480–489, 2018.

[44] A. L, H. JL, J. AR, T. G, G. KH, and S. R, "Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study," *Dev Med Child Neurol*, vol. 52, no. 8, pp. 773–778, 2010.

[45] 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 *IEEE EMBC*, 2018, pp. 3598–3601.

[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. 42 314–42 324, 2021.

[47] D. Das, K. Fry, and A. M. Howard, "Vision-based detection of simultaneous kicking for identifying movement characteristics of infants at-risk for neurodisorders," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1413–1418.

[48] 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," *Journal of Clinical Medicine*, vol. 9, no. 1, 2020.

[49] M. Zhu, Q. Men, E. Ho, and et al., "A two-stream convolutional network for musculoskeletal and neurological disorders prediction," *J. Med. Syst.*, vol. 46, no. 76, 2022.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[51] T. Luo and et al., "Weakly supervised online action detection for infant general movements," in *MICCAI*, 2022.

[52] H. Rahmati, O. Aamo, O. Stavdahl, R. Dragon, and L. Adde, "Video-based early cerebral palsy prediction using motion segmentation," *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, p. 3779–3783, 2014.

[53] R. Mills, D. Levac, and H. Sveistrup, "Kinematics and postural muscular activity during continuous oscillating platform movement in children and adolescents with cerebral palsy," *Gait and Posture*, vol. 66, pp. 13–20, 2018.

[54] 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.

[55] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[56] M. Lu and et al., "Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity," in *MICCAI*, 2020, pp. 637–647.

[57] L. Aversano, M. L. Bernardi, M. Cimitile, and R. Pecori, "Early detection of parkinson disease using deep neural networks on gait dynamics," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[58] A. S. Alharthi, A. J. Casson, and K. B. Ozanyan, "Gait spatiotemporal signal analysis for parkinson's disease detection and severity rating," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1838–1848, 2021.

[59] S. Alle and U. D. Priyakumar, "Linear prediction residual for efficient diagnosis of parkinson's disease from gait," in *MICCAI*, 2021.

[60] H. B. Kim, W. W. Lee, A. Kim, H. J. Lee, H. Y. Park, H. S. Jeon, S. K. Kim, B. Jeon, and K. S. Park, "Wrist sensor-based tremor severity quantification in parkinson's disease using convolutional neural network," *Computers in Biology and Medicine*, vol. 95, pp. 140–146, 2018.

[61] A. Zhang, R. San-Segundo, S. Panev, G. Tabor, K. Stebbins, A. Whitford, F. De la Torre, and J. Hodgins, "Automated tremor detection in parkinson's disease using accelerometer signals," in *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2018, pp. 13–14.

[62] A. B. Oktay and A. Kocer, "Differential diagnosis of parkinson and essential tremor with convolutional lstm networks," *Biomedical Signal Processing and Control*, vol. 56, p. 101683, 2020.

[63] M. H. Li and et al., "Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation," *JNER*, vol. 15, no. 1, pp. 1–13, 2018.

[64] Y. Liu and et al., "Vision-based method for automatic quantification of parkinsonian bradykinesia," *TNSRE*, vol. 27, no. 10, pp. 1952–1961, 2019.

[65] S. Williams and et al., "The discerning eye of computer vision: Can it measure parkinson's finger tap bradykinesia?" *J. Neurol. Sci.*, vol. 416, p. 117003, 2020.

[66] A. Sabo and et al., "Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data," *JNER*, vol. 17, no. 1, pp. 1–10, 2020.

[67] S. Rupprechter and et al., "A clinically interpretable computer-vision based method for quantifying gait in parkinson's disease," *Sensors*, vol. 21, no. 16, p. 5437, 2021.

[68] K. Abe and et al., "Openpose-based gait analysis system for parkinson's disease patients from arm swing data," in *ICAMechS*, 2021, pp. 61–65.

[69] R. Guo and et al., "Sparse adaptive graph convolutional network for leg agility assessment in parkinson's disease," *TNSRE*, vol. 28, no. 12, pp. 2837–2848, 2020.

[70] F. Yang and et al., "Make skeleton-based action recognition model smaller, faster and better," in *ACM multimedia asia*, 2019, pp. 1–6.

[71] T.-Y. Lin and et al., "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[72] M. Lu, K. Poston, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, K. M. Pohl, J. C. Niebles, and E. Adeli, "Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity," in *Med Image Comput Comput Assist Interv (MICCAI)*, 2020.

[73] M. Lu, Q. Zhao, K. L. Poston, Sullivan *et al.*, "Quantifying parkinson's disease motor severity under uncertainty using mds-updrs videos," *Medical Image Analysis*, vol. 73, 2021.

[74] B. Lin and et al., "Bradykinesia recognition in parkinson's disease via single rgb video," *TKDD*, vol. 14, no. 2, pp. 1–19, 2020.

[75] X. Wang, S. Garg, S. N. Tran *et al.*, "Hand tremor detection in videos with cluttered background using neural network based approaches," *Health Inf Sci Syst.*, vol. 9, p. 30, 2021.

[76] R. Guo and et al., "A tree-structure-guided graph convolutional network with contrastive learning for the assessment of parkinsonian hand movements," *MIA*, vol. 81, p. 102560, 2022.

[77] Z. Yin and et al., "Assessment of parkinson's disease severity from videos using deep architectures," *JBHI*, vol. 26, no. 3, pp. 1164–1176, 2021.

[78] W. Liu and et al., "Vision-based estimation of mds-updrs scores for quantifying parkinson's disease tremor severity," *MIA*, p. 102754, 2023.

[79] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers in Neuroscience*, vol. 14, 2020.

[80] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early detection of parkinson's disease using deep learning and machine learning," *IEEE Access*, vol. 8, pp. 147 635–147 646, 2020.

[81] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: A deep learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1618–1630, 2019.

[82] J. M. Hausdorff, "Gait dynamics in parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling," *Chaos (Woodbury, N.Y.)*, vol. 19, no. 2, p. 026113, 2009.

[83] P. Rizek, N. Kumar, and M. S. Jog, "An update on the diagnosis and treatment of parkinson disease," *CMAJ : Canadian Medical Association journal*, vol. 188, no. 16, pp. 1157–1165, 2016.

[84] J. M. Beitz, "Parkinson's disease: A review," *Front Biosci (Schol Ed)*, vol. 6, pp. 65–74, 2014.

[85] J. Pasquini, R. Ceravolo, Z. Qamhawi, J. Lee, G. Deuschl, D. J. Brooks, U. Bonuccelli, and N. Pavese, "Progression of tremor in early stages of parkinson's disease: a clinical and neuroimaging study," *Brain*, vol. 141, pp. 811–821, 2018.

[86] M. D. Hssayeni, J. Jimenez-Shahed, M. A. Burack, and B. Ghoraani, "Wearable sensors for estimation of parkinsonian tremor severity during free body movements," *Sensors (Basel, Switzerland)*, vol. 19, no. 19, p. 4215, 2019.

[87] M.-H. Guo and et al., "Attention mechanisms in computer vision: A survey," *CVM*, vol. 8, no. 3, pp. 331–368, 2022.

[88] V. Mnih and et al., "Recurrent models of visual attention," *NeurIPS*, vol. 27, 2014.

[89] M. Jaderberg and et al., "Spatial transformer networks," *NeurIPS*, vol. 28, 2015.

[90] J. Hu and et al., "Gather-excite: Exploiting feature context in convolutional neural networks," *NeurIPS*, vol. 31, 2018.

[91] X. Wang and et al., "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[92] P. Veličković and et al., "Graph attention networks," in *ICLR*, 2017.

[93] J. Xie and et al., "Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition," *Neurocomputing*, vol. 440, pp. 230–239, 2021.

[94] C. Plizzari and et al., "Skeleton-based action recognition via spatial and temporal transformer networks," *CVIU*, vol. 208, p. 103219, 2021.

[95] J. Zhang and et al., "A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition," *CAAI Trans. Intell. Technol.*, vol. 7, no. 1, pp. 46–55, 2022.

[96] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[97] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[98] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 663–10 671.

[99] J. Chen, H. Sasaki, H. Lai, Y. Su, J. Liu, Y. Wu, A. Zhovmer, C. A. Combs, I. Rey-Suarez, H.-Y. Chang *et al.*, "Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes," *Nature methods*, vol. 18, no. 6, pp. 678–687, 2021.

[100] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[101] A. Vaswani and et al., "Attention is all you need," 2017.

[102] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 68–80.

[103] H. Wang and et al., "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," *arXiv preprint arXiv:2003.07853*, 2020.

[104] A. Dosovitskiy and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[105] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, and N. Jacobs, "Advit: Vision transformer on multi-modality pet images for alzheimer disease diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–4.

[106] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[107] N. A. Asif, Y. Sarker, R. K. Chakrabortty, M. J. Ryan, M. H. Ahamed, D. K. Saha, F. R. Badal, S. K. Das, M. F. Ali, S. I. Moyeen *et al.*, "Graph neural network: A comprehensive review on non-euclidean space," *IEEE Access*, vol. 9, pp. 60 588–60 606, 2021.

[108] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen, "Graph convolutional networks with markov random field reasoning for social spammer detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 1054–1061.

[109] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia, "Learning to simulate complex physics with graph networks," in *International conference on machine learning*, 2020, pp. 8459–8468.

[110] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *The world wide web conference*, 2019, pp. 3307–3313.

[111] E. Khalil, H. Dai, Y. Zhang, B. Dilkina, and L. Song, "Learning combinatorial optimization algorithms over graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[112] A. Sperduti and A. Starita, "Supervised neural networks for the classification of structures," *IEEE transactions on neural networks*, vol. 8, no. 3, pp. 714–735, 1997.

[113] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[114] A. Micheli, "Neural network for graphs: A contextual constructive approach," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 498–511, 2009.

[115] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[116] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE signal processing magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[117] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *ICLR*, 2014.

[118] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[119] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[120] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

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

[122] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[123] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[124] H. Ci, X. Ma, C. Wang, and Y. Wang, "Locally connected network for monocular 3d human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1429–1442, 2020.

[125] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," p. arXiv:1801.07455, 2018.

[126] L. Shi and et al., "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 12 026–12 035.

[127] H. Zhang and et al., "Pose-based tremor classification for parkinson's disease diagnosis from video," in *MICCAI*, 2022.

[128] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature communications*, vol. 11, no. 1, p. 3923, 2020.

[129] H. Singh, G. D. Schiff, M. L. Graber, I. Onakpoya, and M. J. Thompson, "The global burden of diagnostic errors in primary care," *BMJ quality & safety*, vol. 26, no. 6, pp. 484–494, 2017.

[130] M. L. Graber, "The incidence of diagnostic error in medicine," *BMJ quality & safety*, vol. 22, no. Suppl 2, pp. ii21–ii27, 2013.

[131] H. Liang, B. Y. Tsui, H. Ni, C. C. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen *et al.*, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature medicine*, vol. 25, no. 3, pp. 433–438, 2019.

[132] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[133] L. Zhang, Y. Zhao, T. Che, S. Li, and X. Wang, "Graph neural networks for image-guided disease diagnosis: A review," *iRADIOLOGY*, vol. 1, no. 2, pp. 151–166, 2023.

[134] B. Horwitz, "The elusive concept of brain connectivity," *Neuroimage*, vol. 19, no. 2, pp. 466–470, 2003.

[135] B. C. Van Wijk, C. J. Stam, and A. Daffertshofer, "Comparing brain networks of different size and connectivity density using graph theory," *PloS one*, vol. 5, no. 10, p. e13701, 2010.

[136] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, vol. 48, pp. 117–130, 2018.

[137] X. Tang and et al., "A causality-driven graph convolutional network for postural abnormality diagnosis in parkinsonians," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2023.

[138] Z. Sun, H. Yin, H. Chen, T. Chen, L. Cui, and F. Yang, "Disease prediction via graph neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 818–826, 2020.

[139] H. Lu and S. Uddin, "A weighted patient network-based framework for predicting chronic diseases using graph neural networks," *Scientific reports*, vol. 11, no. 1, p. 22607, 2021.

[140] S. N. Golmaei and X. Luo, "Deepnote-gnn: predicting hospital readmission using clinical notes and patient network," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021, pp. 1–9.

[141] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[142] D. Graham, S. 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.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

[147] 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.

[148] H.-S. Fang and et al., "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE PAMI*, vol. 45, no. 6, pp. 7157–7173, 2023.

[149] 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.

[150] 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.

[151] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[152] C. Aizawa, C. Einspieler, F. Genovesi, S. Ibidi, and R. Hause, "The general movement checklist: A guide to the assessment of general movements during preterm and term age," *Jornal de Pediatria*, vol. 97, no. 4, pp. 445–452, 2021.

[153] M. Sadowska, B. Sarecka-Hujar, and I. Kopyta, "Cerebral palsy: Current opinions on definition, epidemiology, risk factors, classification and treatment options," *Neuropsychiatric Disease and Treatment*, vol. 16, pp. 1505–1518, 2020.

[154] I. Tolstikhin and et al., "MLP-mixer: An all-MLP architecture for vision," in *NeurIPS*, 2021.

[155] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *IEEE International Conference on Information Fusion*, 2020, pp. 1–6.

[156] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," in *NeurIPS*, 2022.

[157] A. Karpathy and et al., "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[158] N. Kato, T. Li, K. Nishino, and Y. Uchida, "Improving multi-person pose estimation using label correction," 2018.

[159] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *CVPR*, 2017, pp. 2980–2988.

[160] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014, pp. 568–576.

[161] V. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear svm: a review," *Artif Intell Rev*, vol. 52, p. 803–855, 2019.

[162] Z. Chang, G. A. Koulieris, and H. P. H. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv:2306.04542*, 2023.

[163] G. M. Harshvardhan, K. G. Mahendra, P. Manjusha, and S. R. Siddharth, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, vol. 38, p. 100285, 2020.

[164] W. Liu, X. Lin, X. Chen, Q. Wang, X. Wang, B. Yang, N. Cai, R. Chen, G. Chen, and Y. Lin, "Vision-based estimation of mds-updrs scores for quantifying parkinson's disease tremor severity," *MIA*, vol. 85, p. 102754, 2023.

[165] J. J. Condon and M. R. Scott, "Essential radio astronomy," *Princeton University Press*, 2016.

[166] A. Delval and et al., "Freezing/festination during motor tasks in early-stage parkinson's disease: A prospective study," *Movement Disorders*, vol. 31, no. 12, pp. 1837–1845, 2016.

[167] J. Wang, S. Yan, Y. Xiong, and D. Lin, "Motion guided 3d pose estimation from videos," in *ECCV*, 2020, pp. 764–780.

[168] S. Sveinbjornsdottir, "The clinical symptoms of parkinson's disease," *Journal of neurochemistry*, vol. 139, pp. 318–324, 2016.

[169] S. Li, Z. Gao, and H. Lin, "Lookhops: light multi-order convolution and pooling for graph classification," *arXiv preprint arXiv:2012.15741*, 2020.

[170] S. Fahn, "Description of parkinson's disease as a clinical syndrome," *Annals of the New York Academy of Sciences*, vol. 991, pp. 1–14, 2003.

[171] Z. Gao and et al., "Lookhops: light multi-order convolution and pooling for graph classification," *arXiv preprint arXiv:2012.15741*, 2020.

[172] S. Fahn, "Description of parkinson's disease as a clinical syndrome," *Annals of the New York Academy of Sciences*, vol. 991, no. 1, pp. 1–14, 2003.

[173] S. L. Pintea and et al., "Hand-tremor frequency estimation in videos," in *ECCV Workshops*, 2018, pp. 0–0.

[174] P. Bain and et al., "Assessing tremor severity," *JNNP*, 1993.

[175] M. Lu and et al., "Quantifying parkinson's disease motor severity under uncertainty using mds-updrs videos," *MIA*, vol. 73, p. 102179, 2021.

[176] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei, "Retentive network: A successor to transformer for large language models," *arXiv preprint arXiv:2307.08621*, 2023.

[177] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[178] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[179] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[180] Y. Wang, Y. Xu, J. Yang, M. Wu, X. Li, L. Xie, and Z. Chen, "Fully-connected spatial-temporal graph for multivariate time-series data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 715–15 724.

[181] K. Yi, Q. Zhang, W. Fan, H. He, L. Hu, P. Wang, N. An, L. Cao, and Z. Niu, "Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[182] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.