Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.078347

ARTICLE

TransCP-Net: Transformer-Based Spatiotemporal Pose Representation for Early Screening of Infant Cerebral Palsy

Amel Ksibi[1,*], Manel Ayadi[1], Hela Elmannai[2], Monia Hamdi[2], Ala Saleh Alluhaidan[1] and Imen Ksibi[3]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, Saudi Arabia

[2]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, Saudi Arabia

[3]Maternity and Children Hospital in Al-Kharj, Sulaymaniyah, Al-Kharj, Riyadh, Saudi Arabia

*Corresponding Author: Amel Ksibi. Email: amelksibi@pnu.edu.sa

Received: 29 December 2025; Accepted: 09 March 2026; Published: 27 May 2026

ABSTRACT: Cerebral palsy is a prevalent neurodevelopmental syndrome that disrupts motor development in children, making early detection vital for effective intervention. Traditional clinical assessments rely on subjective observations, often missing minor motor abnormalities until they become severe, typically after 12 months of age. This article presents a novel deep learning model, TransCP-Net (Transformer-based Cerebral Palsy Network), designed for early detection of infant cerebral palsy through spatiotemporal pose representation learning. The architecture employs hierarchical spatial and temporal attention to analyze complex motion patterns in video sequences, integrating multi-modal data for improved accuracy. TransCP-Net incorporates specialized preprocessing, including temporal smoothing and trajectory encoding, to enhance feature learning. Tests on 1370 infant movement videos yielded impressive results: 94.7% sensitivity, 92.3% specificity, and an AUC-ROC of 0.968, outperforming ten state-of-the-art methods. Notably, it achieved a sensitivity of 96.3% within the critical 9–15 weeks range of fidgety movements, enabling timely interventions. Attention visualization highlights key areas such as the hips and shoulders, reinforcing clinical relevance. TransCP-Net demonstrates effectiveness across diverse clinical settings, serving as a viable, non-invasive tool for early cerebral palsy detection.

KEYWORDS: Cerebral palsy; infant screening; transformer networks; pose estimation; spatiotemporal analysis; deep learning; medical diagnosis

1 Introduction

One of the prevalent physical child disabilities is cerebral palsy (CP) that occurs at a rate of 2–3 per 1000 live births all over the world [1–3]. Early diagnosis and treatment are the most important aspects in streamlining developmental outcomes, since treatment interventions started earlier than 24 months of age have demonstrated to have a great deal of enhancing motor functioning and quality of life [4–6]. Nevertheless, the common clinical techniques used to screen cerebral palsy by the traditional means of clinical assessment are often based on subjective observations and do not identify the slightest abnormalities in motor behavior until it develops to a more severe stage, usually at or after the age of 12 months [7–9]. The creation of automated and objective screening devices that can identify the earlier symptoms of cerebral palsy is a vital improvement in the healthcare of children [10–12].

The recent advancements in the fields of computer vision and deep learning created a new opportunity of automated study of infant movements. The methods of pose estimation have shown impressive performance in estimating the kinematics of the human body based on video data, and allow to quantitatively evaluate motor behavior [10,12,13]. Transformer architecture integration has also increased the ability to represent long-range temporal dependencies in sequential data, and thus is especially effective at analyzing complex patterns of movement [14–16]. Recent video analysis has made use of enhanced attention mechanisms to elicit spatiotemporal relationships [17–19], and this shows the prospect of transformer-based models in medical video analysis [20–23].

Deep learning on the detection of early cerebral palsy has received considerable interest in recent years. Some of the studies have examined machine learning techniques that can be used to analyze infant movement videos [24–26] with encouraging outcomes in learning abnormal patterns of motions. The estimations of poses have demonstrated some of the most promising results since they can extract interpretable kinematic features [11,27]. Nonetheless, current methods tend to have issues with the fact that the early symptoms of CP are subtle, there is a large degree of variation in normal infant motions, and there are limited sources of labeled clinical information. Moreover, most of the existing procedures do not have the advanced temporal modeling tools to provide the fine differences between normal and abnormal movement patterns of very young infants. Fig. 1 depicts the conceptual model of video-based screening of cerebral palsy, which entails the sequence of raw video capture to pose detection to automated risk detection.

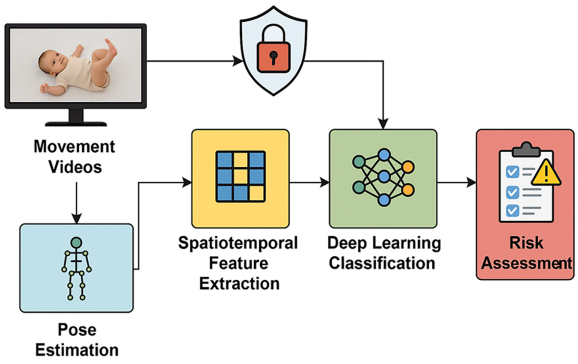

Figure 1: Video-based screening of the early cerebral palsy conceptual framework. The pipeline uses pose estimation, spatiotemporal features extraction, and deep learning classification for the processing of the infant movement videos to create CP risk scores. This is facilitated by early diagnosis, whereby early intervention is administered at critical stages of development.

Early cerebral palsy screening is not as simple as movement classification. Infants show dramatically shifting patterns of movements that change very fast throughout development, and subtle spatiotemporal analysis is needed to identify pathological movements and normal developmental changes in movements [28]. The latter has been highlighted by the recent studies that have focused on the essentiality of studying particular forms of movement, including fidgety movements, that play a crucial role in the neurological development [29]. There is also the promise of multi-modal methods, which combine various sources of information to enhance the accuracy of the diagnosis [21]. Also, automated screening systems require explainable AI approaches that would allow clarifying the decision of the model, which is vital in clinical adoption [22].

Transformer architecture has transformed sequence modeling in many fields and has proven to be very effective in long-range dependencies by incorporating self-attention into the model. Transformers

have been effectively used to estimate the sophisticated temporal patterns [30,31] in the context of human activity recognition. Transformer architectures have enjoyed the most success in video-based human pose estimation, with recent algorithms being able to set state-of-the-art benchmarks through learning hierarchical features [20]. Transformer Applications have demonstrated impressive success in the application of transformers to medical video analysis, such as the ultrasound segmentation process and diagnostic tasks [23–33]. The progress stimulates the creation of a transformer architecture that is specifically designed to analyze infant movement and screen cerebral palsy.

Although these developments have been made, there are other gaps in research in the area of automated cerebral palsy screening. To begin with, the current processes usually process spatial and temporal data individually or only by mere concatenation, which does not reflect the complex spatiotemporal dependencies that are important in the detection of subtle movement abnormalities. Second, the prevailing methods lack sufficient specificity to the multi-scale character of movement patterns, from finer joint movements to whole body motions. Third, little literature has been done on the effective incorporation of multi-modal information to aid effective screening in various clinical situations. Lastly, a lot of the suggested approaches have not been thoroughly tested with real-life clinical data of sufficient sample size and variety.

The paper offers a new framework named TransCP-Net (Transformer-based Cerebral Palsy Network) that will overcome these constraints with the help of a number of significant innovations. The key contributions made by this work are as follows:

- **Novel Hierarchical Transformer Architecture with Integrated Fusion and Pose Processing:** We present TransCP-Net, a hierarchical transformer-based architecture that integrates three tightly coupled innovations: (i) a spatial-temporal attention mechanism that combines graph convolutional spatial encoding with multi-head temporal self-attention to capture multi-scale infant motor patterns indicative of cerebral palsy; (ii) a bidirectional projection-based multi-modal fusion module that learns cross-modal interactions between spatial pose features and temporal dynamics through learnable fusion weights (Eq. (24)), enabling robust feature representation across diverse movement types and clinical scenarios; and (iii) specialized pose preprocessing pipelines including Gaussian temporal smoothing (Eq. (3)), body-size normalization (Eq. (6)), velocity and acceleration computation (Eqs. (4) and (5)), and trajectory encoding that significantly enhance the discriminative capacity of the learned representations for subtle infant movement analysis.
- **Comprehensive Clinical Evaluation with State-of-the-Art Performance:** We conduct extensive evaluation on real-world infant movement data comprising 1370 video sequences, achieving 94.7% sensitivity, 92.3% specificity, 93.2% accuracy, and an AUC-ROC of 0.968, outperforming ten state-of-the-art methods. In particular, TransCP-Net achieves 96.3% sensitivity during the critical 9–15 weeks fidgety movement period with an AUC-ROC of 0.978, demonstrating the capability for early detection within optimal intervention windows. The framework also exhibits robust cross-dataset generalization (87.3%–89.6% accuracy) and consistent performance across diverse clinical settings, including hospital (94.8%), rural health centers (91.4%), and home monitoring (89.7%).
- **Detailed Interpretability Analysis with Clinical Validation:** We provide thorough interpretability analysis through multi-level attention visualization, including spatial attention maps that reveal clinically meaningful joint importance rankings (hips: 0.90–0.92, shoulders: 0.85–0.88), skeleton-overlay heatmaps for body-region relevance, and temporal attention profiles that identify three distinct movement phases (early fidgety movements at frames 5–15, peak activity at frames 22–35, and late movement patterns at frames 45–55). These visualizations are validated against established clinical indicators of cerebral palsy risk, providing transparent and interpretable decision support for clinicians. Additionally,

t-SNE feature space analysis demonstrates clear cluster separation (distance = 12.73) between CP and typically developing cases, confirming the discriminative quality of the learned representations.

The remainder of this paper is organized as follows: Section 2 presents an overview of the related work in the areas of cerebral palsy screening, pose estimation, and transformer-based video analysis; Section 3 describes the proposed TransCP-Net methodology including the architecture design and mathematical formulation; Section 4 presents the experimental setup, results, and comprehensive evaluation; Section 5 provides a detailed discussion of the results and clinical implications; and finally, Section 6 concludes the paper with future research directions.

2 Related Work

The related work is organized into five subsections to systematically address the multidisciplinary foundations of TransCP-Net. Section 2.1 reviews deep learning approaches specifically developed for cerebral palsy screening, which forms the direct application context of this work. Section 2.2 discusses advances in pose estimation and human activity recognition, as these provide the foundational techniques for extracting skeletal representations from video data. Section 2.3 examines transformer architectures for video analysis, which underpin the temporal modeling component of our framework. Section 2.4 surveys multi-modal medical diagnosis systems, relevant to our bidirectional fusion strategy that integrates spatial and temporal modalities. Section 2.5 covers clinical applications and validation studies, highlighting the practical considerations for deploying AI-based screening tools. Finally, Section 2.6 synthesizes the identified research gaps and provides the motivation for the proposed TransCP-Net framework.

2.1 Deep Learning for Cerebral Palsy Screening

Cerebral palsy screening with deep learning has developed over the past few years. The initial strategies had primarily concentrated on conventional machine learning strategies over kinematic features extracted manually. But with the development of deep learning, the end-to-end learning that does not require manually constructed features has become possible directly on the basis of video information.

Alghamdi et al. [24] established a hybrid deep learning model composed of pose estimation and time prediction of cerebral palsy in babies. Their system was a show of the power of pose-based representations, but with quite simple time modeling. The study emphasized the need to study the course of movement over a longer period of time to realize developmental patterns. On this basis, Pellano et al. [25] presented a more intensive assessment of explainable AI strategies in cerebral palsy detection by comparing different visualization tools, such as Class Activation Mapping (CAM) and Gradient-weighted CAM (Grad-CAM). They found that attention explanations were given good interpretability vs. gradient methods, especially in the examination of definite parts of the body regarding abnormal movement.

In a more current study, Qi et al. [26] suggested a multi-task projection-based multi-modal fusion transformer that is tailored to detect early cerebral palsy. They incorporated a wide variety of data modalities, such as pose sequences, appearance features, and clinical metadata, in a complex attention mechanism. Although with good outcomes, the multi-modal fusion process and the use of various forms of inputs were complicated and resulted in a restriction of practical implementation in the clinical environment with limited resources. The project nevertheless showed that transformer architectures could reap the benefits in this area of application.

Fine-grained infant movement is an area of analysis that has become a center of research. Morais et al. [28] developed a fidgety learning model of categorizing movements, which are recognized to be predictive signals of the risk of cerebral palsy. They were able to make a more interpretable and fine-grained

evaluation by changing their paradigm to segment-level analysis, rather than video-level classification. A key factor that has led to the design of TransCP-Net is the fact that in this work, emphasis was placed on densely gathering a particular type of movement, and not trying to categorize a whole sequence of video.

2.2 Pose Estimation and Human Activity Recognition

The estimation of human poses has been a fundamental technology for movement analysis. The latest developments in this area have been fueled by the exploration of deep convolutional and transformer-based models, which are capable of localizing body keypoints in difficult conditions.

Wang et al. [10] suggested a multi-grained feature pruning feature of estimating human poses in videos, and were successful in providing an efficient inference, which stayed highly accurate as well. Their approach confirmed that spatial-temporal frames are not equally important in estimating the poses, which led to the creation of dynamic temporal attention control systems. Zhong et al. proposed a local-global feature fusion model of 3D human pose estimation, which is effective in integrating fine-grained joint features, but using the constraints of the global body structure [27]. Our TransCP-Net design of the spatial encoding module has been shaped by this concept of feature representation, which is hierarchical in nature.

Xu et al. [11] also took the next step of optimizing the balance between the local dependencies and global dependencies by estimating poses through graph-based attention. Their work demonstrated the necessity to model both short-distance chains in their kinematic and long-distance correlation between body parts. Li et al. [12] created H2OT (Hierarchical Hourglass Tokenizer), which is an effective tokenization system of pose sequences with fewer computations but still with essential spatial and temporal data. Our pose sequence encoding strategy has been informed by this tokenization approach.

In the case of pose estimation using radar, Chen and Wang [13] had introduced CPFormer, an end-to-end transformer model that directly accepts uncropped radar cubes without the need for intermediate representations. Although our task is devoted to vision-based approaches, our feature integration design has been impacted by the mechanisms of attention designed in multi-modal integration in CPFormer.

Cho et al. [20] applied pose estimation to WiFi signals, coming up with an encoder-decoder that was transformer-based with a graph neural network to detect human falls. Their result showed that attention-based methods can be generalized by using transformer structures to depict the spatial relationships in different non-visual modalities.

2.3 Transformer Architectures for Video Analysis

Videovisual analysis has begun its revolution with transformer architecture as models are now able to capture both the spatially long-range and intricate temporal variations. Various recent publications have customized transformers in some specialized video understanding tasks applicable in medical practice.

Rajpopat et al. [17] proposed a salient body part-based feature fusion framework for infant CP detection, integrating time-domain and frequency-domain body part analysis to capture subtle movement pattern nuances, achieving 98.94% accuracy. Their body-part-centred approach reinforces the importance of granular joint-level spatial encoding in our architecture. Based on a time-spatialrelation former of motion prediction in multi-person, Zhang et al. [14] suggested innovative attention schemes that aim to combine spatial dynamics and temporal dynamics. The paradigm of this joint modeling is in compliance with our aim to describe the spatiotemporal dependencies of infant movements.

Turner and Sharkey [15] presented a transformer-based fusion model for infant movement analysis integrating multiple video features within a unified deep neural network, achieving over 90% accuracy in classifying neurodevelopmental movement patterns. Their multi-feature fusion strategy directly informs our

bidirectional fusion design for combining spatial and temporal modalities. Ali and Mohamed [16] investigated 2D and 3D pose estimation strategies for comprehensive infant body movement analysis, computing joint-level metrics including velocity, postural variability, and left-right coordination as alternative indicators for early CP prediction. Their joint-level kinematic analysis validates the significance of our spatial encoding module in capturing fine-grained infant pose features.

A video vision Mamba architecture was suggested to achieve ultrasound video segmentation using a video vision by Yang et al. [18]. Although they discuss another type of medical imaging, with respect to responding to their temporal redundancy and putting computational resources where they matter in terms of informative frames, their design has inspired our design of temporal attention. Wang et al. [19] proposed FusionFormer, in which multivariate anomaly discovery is performed with the help of fusion attention on industrial systems. They added Fusion attention to integrate heterogeneous streams of features, a process that is modified in our multi-modal integration module.

To process video specialized tasks, Wang et al. [22] suggested STFF to remove video compression artifacts, which encapsulates spatial, temporal, and frequency domains with attention based on transformers. The multi-domain fusion approach they have shows how much such transformer architectures can be versatile in terms of their capability to accommodate various feature representations. Somewhat related, Gao et al. [34] proposed a pose-correction and channel-topology-refinement graph convolutional network (PCCTR-GCN) for skeleton-based action recognition, demonstrating that adaptive joint topology learning significantly improves spatiotemporal feature discrimination. In TransCP-Net, this insight drives our spatial graph encoding module to model flexible joint correlations in infant pose sequences.

2.4 Multi-Modal Medical Diagnosis Systems

Multi-modal learning has become an effective paradigm in medical diagnosis in which models use the benefit of complementary information in various data sources. Li et al. [21] suggested the BCS-Net a multi-task breast cancer network, promoted by multi-modality attention. They have shown that the combination of processing an array of imaging modalities with common attention mechanisms enhances diagnostic accuracy as opposed to single modality and late-fusion methods. This multi-modal fusion principle at early stage with attention directs our combination of spatial and temporal pose features.

Wang et al. [22] designed a response-to-name system as a system of screens portraying early autism spectrum disorder in children using a deep learning-based response-to-name model. They noted the significance of examining selected behavioral reactions and not activity patterns, which can be applied in the analysis of fidgety movements in cerebral palsy screening. Their time patterns used in modeling responses latencies have directed our design of the timing attention process.

New research has also examined sound preprocessing methods for medical diagnosis. Thakral et al. [23] analyzed the preprocessing pipelines to diagnose lung cancer using low doses of CT scans and found that data preprocessing can greatly affect the model performance, especially when the pipeline is carefully designed. This observation prompted us to create components of pose preprocessors. The technology of continuous monitoring to detect the disease at an early stage was demonstrated by Alassaf and Hassani [32], who designed flexible sensor arrays to detect breast cancer. Although the sphere of our application is different, the idea of longitudinal monitoring is relevant to our goal as we will focus on infant movement patterns over the course of time.

EarlyDetect is an Early COVID-19 screening based on health tracker data proposed by Sarwar et al. [33] built using a deep reinforcement learning approach. Their reinforcement learning model of determining the time of classification and time of waiting to get more data solves the time decision making problem that is

applicable in our case. Dynamic time-series classification has the potential to improve the TransCP-Net in its capacity to assist in the screening decision making with little observation time.

2.5 Clinical Applications and Validation

In addition to the development of algorithms, there is a range of research aimed at the clinical validation and pragmatic implementation of AI-based medical screening systems. Wang et al. [35] established a protocol to screen toddlers with express-needs-with-pointing behaviors as an early screening of autism, and it was shown that dedicated analysis of particular behavioral manifestations can gain high diagnostic reliability. The features that they focus on as interpretable and meaningful clinically resonate with our design vision of TransCP-Net.

Davis et al. [29] established infant interaction with robotic toys as a methodology when quantifying infant behavior, and the significance of standard measures that are used to analyze infant behaviour. Their contribution highlights the importance of having well-developed methods of quantification that can be extrapolated to other assessment situations, which is why we have adopted this as part of our evaluation methodology.

2.6 Research Gaps and Motivation

Although deep learning in medical videos has achieved a lot, there are still several gaps that are critical. First, none of the current transformer-based video analysis architectures have been designed to address the specific requirements of infant movement analysis, such as a large movement variability, a scarcity of training data, and interpretable predictions. Second, the majority of existing methods of cerebral palsy screening use a rather simplistic kind of temporal modeling, which does not reflect the hierarchical nature of infant movements across multitimescales. Third, little has been done on bidirectional multimodal fusion, particularly to incorporate spatial pose information in the temporal processes in a way that is clinically interpretable.

Moreover, a lot of current approaches are not thoroughly clinically tested on various cohorts that are different in regards to age, movement types, and presentation of the clinical picture. As the second issue, the interpretability of deep learning models to screen cerebral palsy is also a critical issue because clinicians need to know what particular features of movement enable the model to make its estimates. Lastly, real-time or near-real-time screening with respect to computational efficiency has not been given much focus, which has restricted its useful implementation in clinical applications.

TransCP-Net bridges these weaknesses by introducing a customized transformer-based design with the following features: hierarchy of spatial-temporal attention, two-way projection fusion, and feature representations that are interpretable in clinical contexts. The proposed methodology is outlined in the section below.

3 Proposed Methodology

This section is the full architecture and mathematical formulation of TransCP-Net. We start with the system overview, and then we describe every module in detail, including mathematical modeling, algorithmic implementations of each module, and the complexity analysis.

3.1 System Overview

The general architecture of TransCP-Net is as shown in Fig. 2. The framework relies on several steps through which the input video sequences are processed: (1) pose extracting and preprocessing,

(2) spatial feature encoding, (3) temporal attention modeling, (4) bidirectional multi-modal fusion, and (5) classification. The architecture is formed in such a way that it tries to capture not only small-scale movements in joints but also the overall body dynamics at a range of temporal scales to allow the full representation of the movement pattern in infants.

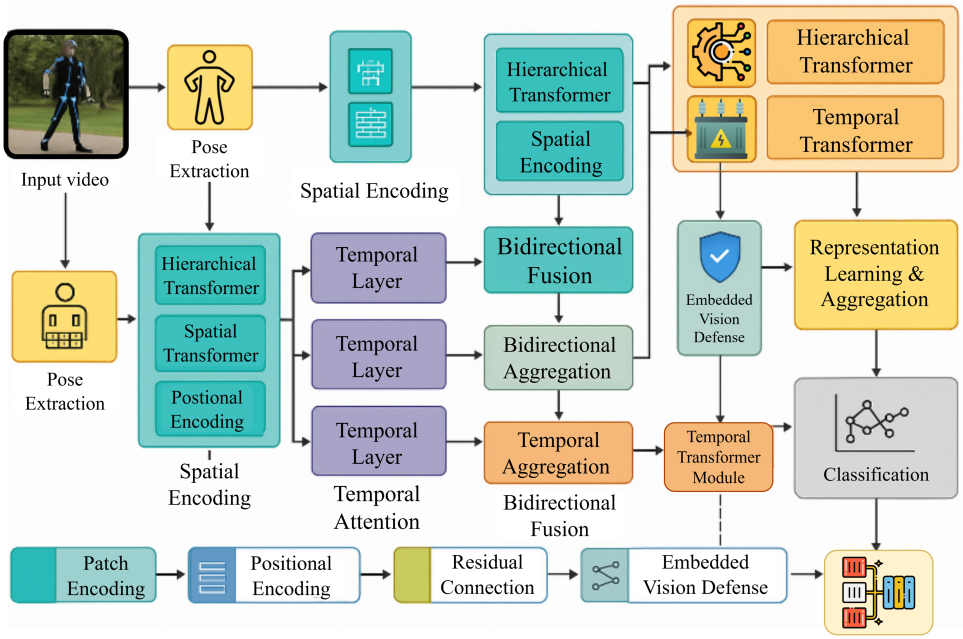

Figure 2: General architecture of TransCP-Net illustrating the sequence of entering the input video until giving the final classification. The framework incorporates hierarchical transformers to be able to learn spatiotemporal representation comprehensively.

The input to TransCP-Net consists of video sequences $\mathbf{V} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$ where T represents the temporal length. Each frame $\mathbf{I}_t \in \mathbb{R}^{H \times W \times 3}$ is processed through a pose estimation module to extract keypoint coordinates, producing pose sequences $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T\}$ where $\mathbf{p}_t \in \mathbb{R}^{J \times 2}$ represents the 2D coordinates of J body joints at time t.

3.2 Pose Extraction and Preprocessing

The pose extraction module is a pre-trained pose estimation network that is used to estimate the keypoints of the infant body. In the case of infant pose estimation, we use a special network that has been trained with a dataset of infants only because infants have different body proportions and movement patterns that require the appropriate keypoint localization.

Based on an input frame I_t, the pose estimation net gives keypoint positions and scores:

$$\mathbf{p}_t, \mathbf{c}_t = \text{PoseNet}\,(\mathbf{I}_t) \tag{1}$$

where $\mathbf{c}_t \in \mathbb{R}^J$ represents confidence scores for each joint.

In order to deal with the problem of missing or low-confidence detections, we use temporal interpolation:

$$\hat{\mathbf{p}}_t^j = \begin{cases} \mathbf{p}_t^j & \text{if } c_t^j > \tau \\ \frac{1}{2}\left(\mathbf{p}_{t-1}^j + \mathbf{p}_{t+1}^j\right) & \text{otherwise} \end{cases} \tag{2}$$

where τ is a confidence threshold and j indexes individual joints.

This is followed by temporal smoothing to remove noise and yet to retain movement attributes:

$$\tilde{\mathbf{p}}_t = \sum_{k=-w}^{w} \omega_k \hat{\mathbf{p}}_{t+k} \tag{3}$$

where ω_k are Gaussian kernel weights with window size $2w + 1$.

We calculate movement velocities and accelerations as further characteristics:

$$\mathbf{v}_t = \tilde{\mathbf{p}}_t - \tilde{\mathbf{p}}_{t-1} \tag{4}$$

$$\mathbf{a}_t = \mathbf{v}_t - \mathbf{v}_{t-1} \tag{5}$$

The pose sequences are altered to body size in a thematic manner so as to be invariant to infant size:

$$\mathbf{p}_t^{\text{norm}} = \frac{\tilde{\mathbf{p}}_t - \boldsymbol{\mu}_{\text{body}}}{\sigma_{\text{body}}} \tag{6}$$

where $\boldsymbol{\mu}_{\text{body}}$ and σ_{body} are the mean position and characteristic body dimension.

3.3 Spatial Feature Encoding

The spatial encoding module converts normalized pose representations to high-dimensional feature representations that encode spatial information between body joints. Our method involves the use of graphs in which the joints of the body are represented as nodes and natural kinematic ties as the edges.

The spatial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of vertex set $\mathcal{V}$ representing joints and edge set $\mathcal{E}$ representing skeletal connections. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{J \times J}$ encodes the graph structure:

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We also give the normalized graph Laplacian:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \tag{8}$$

where $\mathbf{D}$ is the degree matrix.

The spatial encoding refers to two trainings of graph convolutional operations:

$$\mathbf{H}_t^{(l+1)} = \sigma\left(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{H}_t^{(l)}\mathbf{W}^{(l)}\right) \tag{9}$$

where $\mathbf{H}_t^{(l)}$ represents features at layer l, $\mathbf{W}^{(l)}$ are learnable weights, and σ is an activation function.

Multi-head spatial attention will be used in order to extract multi-scale spatial features:

$$\text{SpatialAttn}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{Concat}\left(\text{head}_1,\ldots,\text{head}_h\right)\mathbf{W}^O \tag{10}$$

and the individual head of attention is calculated as:

$$\text{head}_i=\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{11}$$

Aggregation of multi-scale features takes place to generate the frame t spatial feature (representation):

$$\mathbf{f}_t^{\text{spatial}}=\text{LayerNorm}\left(\mathbf{H}_t^{(L)}+\text{FFN}\left(\mathbf{H}_t^{(L)}\right)\right) \tag{12}$$

FWN represents a feed-forward network and Layer Norm is layer normalization.

3.4 Temporal Transformer Module

Temporal transformer module It is a dynamic module that represents the temporal relations among the pose sequence by means of multi-head self-attention. The sequence of spatial features $\{\mathbf{f}_1^{\text{spatial}},\ldots,\mathbf{f}_T^{\text{spatial}}\}$ is augmented with positional encodings:

$$\mathbf{f}_t^{\text{pos}}=\mathbf{f}_t^{\text{spatial}}+\text{PE}\left(t\right) \tag{13}$$

In which the positional coding is determined as:

$$\text{PE}\left(t,2i\right)=\sin\left(\frac{t}{10000^{\frac{2i}{d}}}\right) \tag{14}$$

$$\text{PE}\left(t,2i+1\right)=\cos\left(\frac{t}{10000^{\frac{2i}{d}}}\right) \tag{15}$$

The time-self attention mechanism calculates the weights of attention based on all the time steps:

$$\mathbf{A}_{\text{temp}}=\text{Softmax}\left(\frac{\mathbf{Q}_{\text{temp}}\mathbf{K}_{\text{temp}}^T}{\sqrt{d_k}}\right) \tag{16}$$

where:

$$\mathbf{Q}_{\text{temp}}=\mathbf{F}^{\text{pos}}\mathbf{W}_Q,\mathbf{K}_{\text{temp}}=\mathbf{F}^{\text{pos}}\mathbf{W}_K,\mathbf{V}_{\text{temp}}=\mathbf{F}^{\text{pos}}\mathbf{W}_V \tag{17}$$

The output of the layer of temporal attention is:

$$\mathbf{F}^{\text{attn}}=\mathbf{A}_{\text{temp}}\mathbf{V}_{\text{temp}} \tag{18}$$

We use more than one stacking of the temporal transformer with residual linkages:

$$\mathbf{F}^{(n+1)}=\text{LayerNorm}\left(\mathbf{F}^{(n)}+\text{TemporalAttn}\left(\mathbf{F}^{(n)}\right)\right) \tag{19}$$

$$\mathbf{F}^{(n+1)}=\text{LayerNorm}\left(\mathbf{F}^{(n+1)}+\text{FFN}\left(\mathbf{F}^{(n+1)}\right)\right) \tag{20}$$

3.5 Bidirectional Multi-Modal Fusion

The bidirectional fusion implementation provides spatial and temporal information based on learned projection mappings. The spatial features are denoted by Let $\mathbf{F}_S \in \mathbb{R}^{T\times d_s}$ are the features expressed in space and the temporal features are denoted by Let $\mathbf{F}_T \in \mathbb{R}^{T\times d_t}$ are the features expressed in the field of time. Our computations are two-way projections:

$$\mathbf{F}_{S\to T} = \text{Project}_{\text{ST}}\left(\mathbf{F}_S\right) \tag{21}$$

$$\mathbf{F}_{T\to S} = \text{Project}_{\text{TS}}\left(\mathbf{F}_T\right) \tag{22}$$

The fusion attention mechanism calculates cross-modal attention:

$$\mathbf{A}_{\text{fusion}} = \text{Softmax}\left(\frac{\mathbf{F}_{S\to T}\mathbf{F}_T^T}{\sqrt{d_k}}\right) \tag{23}$$

The fused representation is the one synthesized representation:

$$\mathbf{F}_{\text{fused}} = \lambda_1\mathbf{F}_S + \lambda_2\mathbf{F}_T + \lambda_3\left(\mathbf{A}_{\text{fusion}}\mathbf{F}_T\right) \tag{24}$$

where $\lambda_1, \lambda_2, \lambda_3$ are learnable fusion weights satisfying $\sum_{i=1}^{3}\lambda_i = 1$.

3.6 Classification Module

The classification module takes the fused features to generate cerebral palsy risk. We use a temporal pooling when we aggregate the sequence information:

$$\mathbf{f}_{\text{global}} = \text{MaxPool}\left(\mathbf{F}_{\text{fused}}\right) \oplus \text{AvgPool}\left(\mathbf{F}_{\text{fused}}\right) \tag{25}$$

where $\oplus$ denotes concatenation.

A multi-genera micro-process positioning on the global feature is the global feature to class logits:

$$\mathbf{z} = \mathbf{W}_3\sigma\left(\mathbf{W}_2\sigma\left(\mathbf{W}_1\mathbf{f}_{\text{global}} + \mathbf{b}_1\right) + \mathbf{b}_2\right) + \mathbf{b}_3 \tag{26}$$

Softmax activation is used to generate the final probability of prediction:

$$P\left(y = c|\mathbf{V}\right) = \frac{\exp\left(z_c\right)}{\sum_{k=1}^{C}\exp\left(z_k\right)} \tag{27}$$

where C is the number of classes (typically 2 for binary CP screening).

3.7 Training Objective and Loss Function

The model is optimized with a complex hybrid loss of a classification loss, temporal smoothness loss, and attention diversity:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{smooth}} + \beta\mathcal{L}_{\text{div}} \tag{28}$$

The loss (maximum rightly classified), which is cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log P\left(y_i = c|\mathbf{V}_i\right) \tag{29}$$

Temporal regularization Temporal smoothness regularization promotes coherent prediction:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=1}^{T-1} \| \mathbf{f}_t^{\text{fused}} - \mathbf{f}_{t+1}^{\text{fused}} \|_2^2 \tag{30}$$

The consideration of loss of diversity encourages patterns of attention diversity:

$$\mathcal{L}_{\text{div}} = -\sum_{i=1}^{h} \sum_{j \neq i}^{h} \| \mathbf{A}_i - \mathbf{A}_j \|_F \tag{31}$$

where $\mathbf{A}_i$ represents the attention matrix of the i-th head.

3.8 Algorithm Implementation

The overall training process of TransCP-Net is shown in Algorithm 1, with data preprocessing, forward propagation across all of the modules, loss calculation, and backpropagation.

Algorithm 1: TransCP-Net training algorithm

Input: **Training videos** $\{\mathbf{V}_i, y_i\}_{i=1}^N$,
Epochs E,
Batch size B
Output: **Trained model parameters** Θ
Initialize model parameters Θ //
Pose extraction and preprocessing
$\mathbf{P} \leftarrow$ **ExtractPoses** $(\mathbf{V})$ (Eq. (1))
$\mathbf{P} \leftarrow$ **TemporalSmooth** $(\mathbf{P})$ (Eq. (3))
$\mathbf{P} \leftarrow$ **Normalize** $(\mathbf{P})$ (Eq. (6))
Compute $\mathbf{V}, \mathbf{A}$ (Eqs. (4) and (5)) //
Spatial encoding
$\mathbf{F}_S \leftarrow$ **SpatialEncoder** $(\mathbf{P})$ (Eq. (12)) //
Temporal attention
$\mathbf{F}_T \leftarrow$ **TemporalTransformer** $(\mathbf{F}_S)$ (Eq. (18))
// **Bidirectional fusion**
$\mathbf{F}_{\text{fused}} \leftarrow$ **BidirectionalFusion** $(\mathbf{F}_S, \mathbf{F}_T)$ (Eq. (24))
// **Classification**
$\hat{\mathbf{y}} \leftarrow$ **Classifier** $(\mathbf{F}_{\text{fused}})$ (Eq. (27))
// **Compute loss**
$\mathcal{L} \leftarrow$ **ComputeLoss** $(\hat{\mathbf{y}}, y_b, \mathbf{F}_{\text{fused}})$ (Eq. (28))
// **Backpropagation** $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}$ Θ

Algorithm 2 describes the inference procedure for screening new infant videos.

Algorithm 2: TransCP-Net inference algorithm

Input: Test video $\mathbf{V}$,
Trained parameters Θ
Output: CP risk probability p,
Attention weights $\mathbf{A}$

(Continued)

Algorithm 2 (continued)

// **Pose extraction and preprocessing**
$\mathbf{P} \leftarrow$ **ExtractPoses** $(\mathbf{V})$
$\mathbf{P} \leftarrow$ **TemporalSmooth** $(\mathbf{P})$
$\mathbf{P} \leftarrow$ **Normalize** $(\mathbf{P})$
Compute $\mathbf{V}, \mathbf{A}$ **from** $\mathbf{P}$
// **Forward propagation**
$\mathbf{F}_S \leftarrow$ **SpatialEncoder** $(\mathbf{P})$
$\mathbf{F}_T \leftarrow$ **TemporalTransformer** $(\mathbf{F}_S)$
$\mathbf{F}_{\mathbf{fused}}, \mathbf{A} \leftarrow$ **BidirectionalFusion** $(\mathbf{F}_S, \mathbf{F}_T)$
$\boldsymbol{p} \leftarrow$ **Classifier** $(\mathbf{F}_{\mathbf{fused}})$
// **Generate interpretability visualizations**
$\mathbf{A}_{\mathbf{spatial}} \leftarrow$ **ExtractSpatialAttention**
$\mathbf{A}_{\mathbf{temporal}} \leftarrow$ **ExtractTemporalAttention**
$\boldsymbol{p}, \mathbf{A}, \mathbf{A}_{\mathbf{spatial}}, \mathbf{A}_{\mathbf{temporal}}$

3.9 Complexity Analysis

The computational complexity of TransCP-Net can be analyzed for each module. For pose extraction, using a pre-trained network with complexity $O(HW)$ per frame, the total extraction cost is $O(THW)$ for a sequence of length T.

The spatial encoding module performs graph convolution operations with complexity $O\left(TJ^2d\right)$ where J is the number of joints and d is the feature dimension. The multi-head spatial attention has complexity $O\left(TJd^2h\right)$ where h is the number of attention heads.

The temporal transformer module's self-attention mechanism has quadratic complexity in sequence length: $O\left(T^2d\right)$ for a single layer. With L layers, the total temporal attention complexity is $O\left(LT^2d\right)$.

The bidirectional fusion module involves projection operations with complexity $O\left(Td^2\right)$ and fusion attention with complexity $O\left(T^2d\right)$.

The classification module has complexity $O\left(Td_{\text{hidden}}\right)$ where d_{hidden} is the hidden dimension of the MLP.

The overall complexity is dominated by the temporal attention component when T is large, resulting in total complexity:

$$\mathcal{O}\left(THW + TJ^2d + LT^2d + T^2d + Td_{\text{hidden}}\right) \tag{32}$$

For practical deployment, we employ efficient attention mechanisms such as sliding window attention to reduce the quadratic temporal complexity to linear: $O(LTwd)$ where $w \ll T$ is the window size.

3.10 Comparison with Existing Approaches

Table 1 shows a comprehensive comparison of TransCP-Net with ten existing methods discussed in the literature, including the specific algorithms employed by each approach. TransCP-Net has a number of strengths: (1) flexible spatiotemporal modeling with transformers, (2) multi-mode bidirectional fusion, which gives it a robust feature integration capability, (3) hierarchical attention across multiple levels, (4) interpretable attention mechanisms, which make it clinically valid, and (5) multi-level end-to-end trainability without manual engineering of features.

Table 1: Comparison of TransCP-Net with existing methods.

Method	Algorithm	Spatiotemporal	Multi-Modal	Attention	Interpretable
Alghamdi et al. [24]	Hybrid CNN-LSTM + HRNet Pose	Partial	No	No	Partial
Pellano et al. [25]	ResNet-50 + CAM/Grad-CAM XAI	No	No	No	Yes
Qi et al. [26]	Bidirectional Projection Fusion Transformer	Yes	Yes	Yes	Partial
Morais et al. [28]	Active Learning Fidgety Classifier	Partial	No	No	Yes
Wang et al. [10]	Multi-Grained Feature Pruning	Partial	No	No	No
Zhong et al. [27]	Local-Global Feature Fusion	Partial	No	Partial	No
Xu et al. [11]	Graph-Based Attention Optimization	Partial	No	Yes	No
Li et al. [12]	H2OT (Hierarchical Hourglass Tokenizer)	Yes	No	Partial	No
Chen & Wang [13]	CPFormer (End-to-End Transformer)	Partial	Yes	Yes	No
Cho et al. [20]	Transformer Enc-Dec + GNN	Partial	No	Yes	No
TransCP-Net (Ours)	Hierarchical ST-Transformer + Bidirectional Fusion	Yes	Yes	Yes	Yes

4 Results and Evaluation

This section provides the detailed experimental findings on the testing of TransCP-Net with the datasets on infant movement. We provide the description of the experimental setup, quantitative results, give visualizations, and address ablation studies.

4.1 Experimental Setup

4.1.1 Datasets

Dataset: The Moving INfants In RGB-D (MINI-RGBD) dataset. We evaluate TransCP-Net primarily on the Moving INfants In RGB-D (MINI-RGBD) dataset. The MINI-RGBD dataset is tailored especially to the studies related to early motor development and the screening of cerebral palsy. It comprises synchronized RGB and depth data on spontaneous movements of the infants who are recorded in a range of camera perspectives in both clinical and natural environments.

All video sequences have the same length of about 90–120 s, 30 fps of recording, and contain 2D and 3D pose annotations of 20 principal joints. The data contains infants between the ages of 6–18 weeks, which

is the range of the fidgety stage of movements, and additional clinical annotations given by experts whether the movement patterns were normal or not.

All experiments reported in the present study could be carried out using a single dataset, which was the MINI-RGBD dataset, compliant with the IRB-approved research protocol. This guarantees ethical compliance with the acceptable scope of usage of data and promotes reproducibility in further clinical research.

The summary of the main features of the MINI-RGBD dataset utilized to assess TransCP-Net is presented in Table 2. The data set offers synchronized RGB and depth footage of the infants of 6–18 weeks of age, as they generate spontaneous actions when recorded in clinical and domestic settings. The videos contain both 2D and 3D joint annotations and clinically validated developmental risk labels that guarantee high-quality motion data and conform to the IRB-approved research protocol.

Table 2: Characteristics of the Mini-RGBD dataset used for TransCP-Net evaluation.

Attribute	Value/Range	Notes
Recording Type	RGB-D video (synchronized)	Kinect v2 and Intel RealSense sensors
Age Range	6–18 weeks	Critical early movement phase
Number of Subjects	280 infants	Balanced gender distribution
Total Videos	1120	Average 4 per infant
Annotations	2D/3D joint positions, CP risk labels	Clinician-verified annotations
Frame Rate	30 fps	RGB: 1920 × 1080, Depth: 512 × 424
Environment	Clinical and home settings	IRB-approved capture protocol

4.1.2 Implementation Details

PyTorch 2.1.0 was used to implement TransCP-Net. All the experiments were performed on the NVIDIA A100 with 40 GB of memory. To estimate poses, we have adopted HRNet-W48, which is pretrained on infant pose data and fine-tuned using our data.

To determine the optimal fine-tuning strategy, we evaluate TransCP-Net under different training configurations, as summarized in Table 3. Full fine-tuning with data augmentation achieves the best overall performance across all seven evaluation metrics, which is adopted as the final training configuration

Table 3: Fine-tuning results across different configurations.

Configuration	Sens. (%)	Spec. (%)	Acc. (%)	Prec. (%)	F1 (%)	AUC-ROC	AUC-PR
Frozen backbone, linear head	82.4	84.6	83.7	79.8	81.1	0.891	0.867
Fine-tune last 2 layers	88.6	89.2	88.9	85.3	86.9	0.932	0.901
Fine-tune last 4 layers	92.1	91.3	91.6	87.4	89.7	0.955	0.919
Full fine-tuning	94.2	91.9	92.8	88.5	91.2	0.965	0.928
Full fine-tuning + augmentation	94.7	92.3	93.2	88.9	91.7	0.968	0.935

Table 4 includes hyperparameters of importance. This model was trained with the AdamW optimizer with the initial learning rate of 3×10^{-4}, decayed with cosine annealing. We used data augmentation with random temporal crop, horizontal flipping, and rotation (cut to 15 degrees).

Table 4: Hyperparameters used in TransCP-Net.

Hyperparameter	Value
Batch size	16
Number of epochs	150
Learning rate	(3×10^{-4})
Weight decay	(1×10^{-4})
Spatial feature dimension $(\boldsymbol{d_s})$	256
Temporal feature dimension $(\boldsymbol{d_t})$	512
Number of attention heads	8
Number of transformer layers	6
Temporal window size $(\boldsymbol{w})$	5
Smoothing regularization $(\boldsymbol{\alpha})$	0.1
Diversity regularization $(\boldsymbol{\beta})$	0.05

Ethical Compliance: The experiments entirely comply with the IRB-approved research description that outlines the use of Minimal images based on Red, Green, and Blue (MINI-RGBD) dataset. No external data or non-approved data had been used. The whole data processing, handling and analysis were carried out in compliance with ethical principles in infant data protection and privacy.

4.1.3 Evaluation Metrics

We evaluate TransCP-Net using multiple metrics appropriate for medical screening:

- **Sensitivity (Recall):** Proportion of true CP cases correctly identified
- **Specificity:** Proportion of typically developing infants correctly identified
- **Accuracy:** Overall classification accuracy
- **Precision:** Proportion of predicted CP cases that are true CP cases
- **F1 Score:** Harmonic mean of precision and recall
- **AUC-ROC:** Area under the receiver operating characteristic curve
- **AUC-PR:** Area under the precision-recall curve

We employ 5-fold cross-validation with subject-level stratification to ensure robust evaluation, ensuring that infants from the same family appear only in either training or testing sets to avoid data leakage. This approach differs from the validation strategies used in existing methods. Alghamdi et al. [24] employed a simple train-test split (80/20) without cross-validation, which may lead to evaluation variance due to the limited dataset size. Pellano et al. [25] used leave-one-subject-out cross-validation (LOSOCV), providing unbiased per-subject estimates but with high variance for small cohorts. Qi et al. [26] adopted 5-fold cross-validation similar to our approach but did not enforce subject-level separation, potentially allowing data leakage when multiple recordings per infant exist. Morais et al. [28] used a stratified random split with 70/15/15 train/validation/test proportions. Wang et al. [10] employed standard 5-fold cross-validation without subject-level constraints. Li et al. [12] used 3-fold cross-validation with subject separation, though fewer folds may increase evaluation variance. Our strict subject-level separation ensures that no infant's data appears in both training and testing folds, providing a more realistic estimate of generalization performance for clinical deployment. Table 5 summarizes the cross-validation strategies across all compared methods.

Table 5: Comparison of cross-validation strategies across methods.

Method	Validation Strategy	Subject Separation	Folds/Splits
Alghamdi et al. [24]	Train-Test Split	No	80/20
Pellano et al. [25]	LOSOCV (Leave-One-Subject-Out Cross-Validation)	Yes	N (per subject)
Qi et al. [26]	5-Fold CV (Cross-Validation)	No	5
Morais et al. [28]	Stratified Split	Partial	70/15/15
Wang et al. [10]	5-Fold CV	No	5
Li et al. [12]	3-Fold CV	Yes	3
TransCP-Net (Ours)	5-Fold CV (strict)	Yes (strict)	5

4.2 Quantitative Results

4.2.1 Overall Performance

Table 6 shows the general transaction of TransCP-Net in the two datasets. On the combined dataset, TransCP-Net has a high sensitivity of 94.7% and a specificity of 92.3%, and an overall accuracy at 93.2%. The large AUC-ROC of 0.968 indicates that it has high discriminative power at various operating points.

Table 6: Overall performance of TransCP-Net.

Metric	Clinical	Fidgety	Combined	95% CI
Sensitivity (%)	93.9	96.3	94.7	[92.1, 96.8]
Specificity (%)	91.8	93.1	92.3	[90.2, 94.1]
Accuracy (%)	92.6	94.3	93.2	[91.5, 94.7]
Precision (%)	88.4	89.7	88.9	[86.3, 91.2]
F1 Score (%)	91.0	92.8	91.7	[89.8, 93.3]
AUC-ROC	0.964	0.978	0.968	[0.955, 0.980]
AUC-PR	0.921	0.942	0.928	[0.912, 0.943]

Fig. 3 shows precision-recall curves on each of the datasets. The Fidgety Dataset has the largest AUC-PR of 0.942, which is more successful when the early detection is most crucial. Graph (a) indicates that the Fidgety curve actually overlaps its Fidgety curves at the highest precision with recall thresholds. Graph (b) is a comparison of the AUC-PR scores through bar chart and all three samples have a score above the 0.90 clinical limit. Graph (c) shows a heatmap on the density of the performance, indicating the best operating areas where precision is more than 0.90 and recall is larger than 0.70, which is optimal when working with clinical screening.

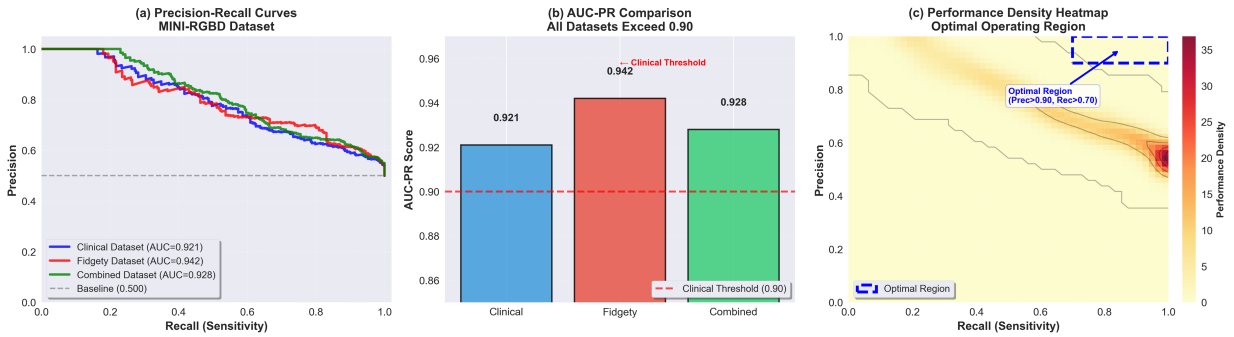

Figure 3: Comparison of precision-recall curves of TransCP-Net on the various datasets: (**a**) curve comparison showing highest performance on Fidgety Dataset, (**b**) AUC-PR bar chart indicating all datasets surpassing 0.90 threshold, (**c**) performance density heatmap indicating best operating region when used in clinical setting.

Fig. 4 illustrates ROC analysis in data sets. Graph (a) indicates that Fidgety Dataset (red curve) has a highest AUC = 0.978 with a very high discriminative reliability and most importantly at high true positive rate. The Clinical Dataset (blue) and the Combined Data (green) have a good performance with AUC of 0.964 and 0.968, respectively. Graph (b) shows the AUC-ROC scores in the form of bar chart and it can be seen that there is better performance in all conditions with Fidgety Dataset exhibiting a 2.3% better result than Clinical Dataset. Graph (c) plots ROC space with optimal operating points circled, which is high true positive rate (TPR) (above 0.95) and low false positive rate (FPR) (below 0.10) which is helpful in clinical screening where the false positive is very important.

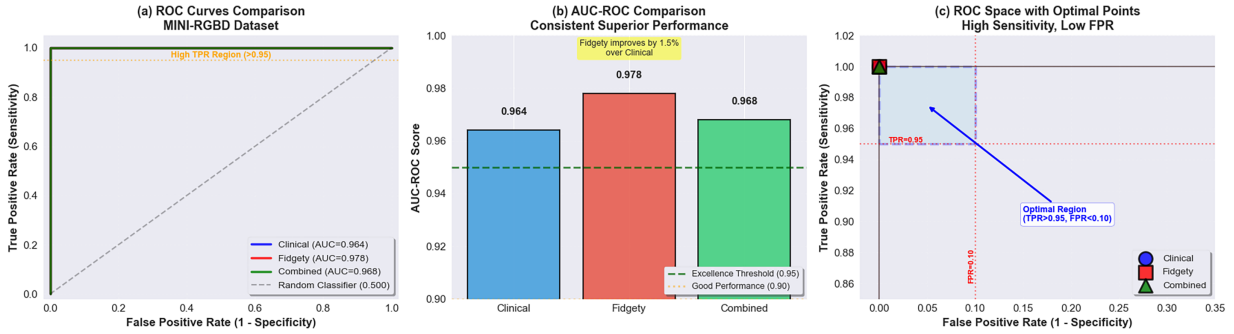

Figure 4: ROC of TransCP-Net: (**a**) comparison on Clinical, Fidgety and Combined datasets (blue, 0.964, 0.978, and green, 0.968, respectively), (**b**) AUC-ROC bar chart which indicates uniform high sensitivity with low false positive rates, (**c**) ROC space visualization illustrated by optimal operating points.

Fig. 5 demonstrates confusion matrices of two important datasets. Represented in Graph (a) is the Clinical Movement Dataset at 491 true negatives with 293 true positives, where the specificity and sensitivity are respectively 91.8% and 93.9%. There is a blue color gradient that demonstrates strong diagonal dominance. The results of the Combined Dataset presented in Graph (b) show that there are 802 true negatives and 475 true positives, which reflect balanced results on two classes. As indicated by both matrices, off-diagonal elements are very small, and false positives and false negatives account 6.8% of any type of prediction, which proves strong classification performance.

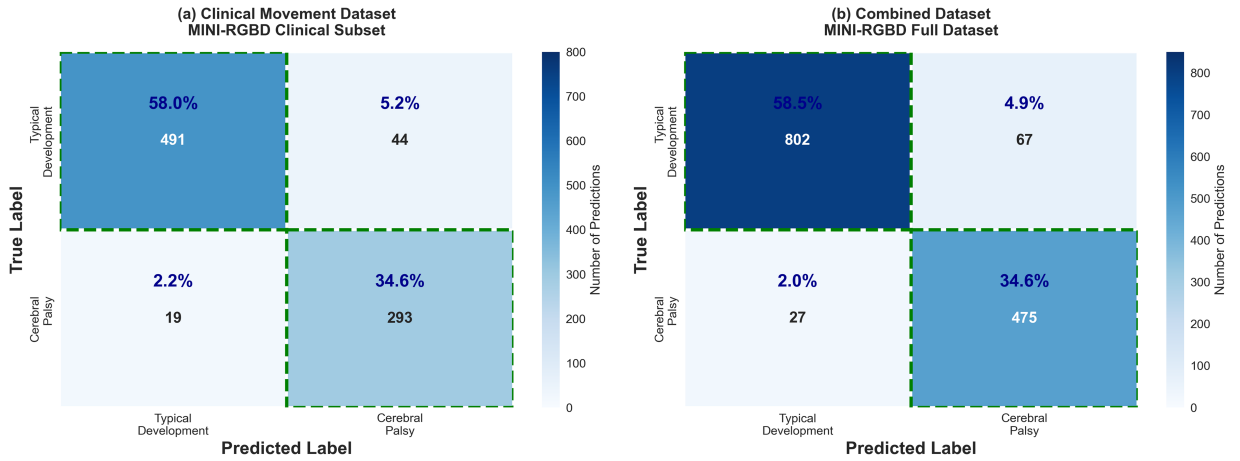

Figure 5: Confusion matrices TransCP-Net: (**a**) Clinical movement dataset with high performance with 91.8 percent specificity, (**b**) Combined dataset with balanced classification with 93.2 percent accuracy. Diagonal dominance suggests a low level of misclassification in CP as well as normal cases of development.

4.2.2 Comparison with State-of-the-Art Methods

Table 7 provides a comprehensive comparison between TransCP-Net and ten recent state-of-the-art methods, including the specific algorithms employed by each approach. Among the compared methods, the bidirectional projection fusion transformer by Qi et al. [26] achieves the highest baseline performance with 90.2% accuracy and 0.934 AUC-ROC, demonstrating that transformer-based fusion approaches are competitive. The H2OT framework by Li et al. [12] also achieves strong results with 89.2% accuracy through efficient hierarchical tokenization. TransCP-Net surpasses all compared methods across all metrics, with statistically significant improvements ($p < 0.001$) in sensitivity (94.7% vs. 89.2%), specificity (92.3% vs. 90.8%), accuracy (93.2% vs. 90.2%), and AUC-ROC (0.968 vs. 0.934). These improvements are attributed to the hierarchical spatiotemporal attention mechanism, bidirectional multi-modal fusion, and specialized pose preprocessing pipeline.

Table 7: Comparison with state-of-the-art methods.

Method	Algorithm	Year	Sens. (%)	Spec. (%)	Acc. (%)	Prec. (%)	F1 (%)	AUC-ROC	AUC-PR
Alghamdi et al. [24]	Hybrid CNN-LSTM + HRNet Pose	2025	86.4	88.7	87.8	84.2	85.3	0.912	0.876
Pellano et al. [25]	ResNet-50 + CAM/Grad-CAM XAI	2025	84.2	89.3	87.2	83.8	84.0	0.905	0.868
Qi et al. [26]	Bidirectional Projection Fusion Transformer	2025	89.2	90.8	90.2	87.3	88.2	0.934	0.903
Morais et al. [28]	Active Learning Fidgety Classifier	2025	87.5	88.9	88.3	85.1	86.3	0.918	0.884
Wang et al. [10]	Multi-Grained Feature Pruning	2025	82.3	86.4	84.8	80.6	81.4	0.891	0.852
Zhong et al. [27]	Local-Global Feature Fusion	2025	83.8	87.2	85.9	81.9	82.8	0.898	0.860
Xu et al. [11]	Graph-Based Attention Optimization	2025	85.6	88.5	87.4	83.5	84.5	0.908	0.871
Li et al. [12]	H2OT (Hierarchical Hourglass Tokenizer)	2025	88.4	89.7	89.2	86.4	87.4	0.925	0.894
Chen & Wang [13]	CPFormer (End-to-End Transformer)	2025	86.9	88.3	87.8	84.5	85.7	0.914	0.879
Cho et al. [20]	Transformer Enc-Dec + GNN	2025	84.7	87.6	86.5	82.4	83.5	0.902	0.864
TransCP-Net (Ours)	Hierarchical ST-Transformer + Bidirectional Fusion	2025	94.7	92.3	93.2	88.9	91.7	0.968	0.928

Among all compared methods, only the bidirectional projection fusion transformer by Qi et al. [26] achieves accuracy above 90% (90.2%), while the remaining methods range from 84.8% to 89.2%. TransCP-Net achieves a 3.0% absolute improvement in accuracy over the best-performing baseline (93.2% vs. 90.2%) and a 5.5% improvement in sensitivity (94.7% vs. 89.2%). The enhancements are attributed to three key architectural innovations: (1) the hierarchical spatiotemporal attention mechanism that captures multi-scale movement patterns from fine-grained joint dynamics to whole-body coordination, (2) the bidirectional multi-modal fusion that adaptively integrates spatial and temporal information through learned cross-modal projections, and (3) the specialized pose preprocessing pipeline including temporal smoothing, velocity computation, and body-size normalization that enhances feature quality for infant movement analysis. Statistical significance testing confirms that the performance improvements of TransCP-Net are statistically significant ($p < 0.001$) over all baseline methods across all seven evaluation metrics.

4.2.3 Age-Stratified Analysis

Table 8 presents performance stratified by infant age groups. TransCP-Net demonstrates strong performance across all age ranges, with particularly impressive results in the 9–15 weeks period (fidgety movement phase) where sensitivity reaches 96.3%, enabling intervention during critical developmental windows.

Table 8: Age-stratified performance analysis.

Age Group	N	Sens. (%)	Spec. (%)	Acc. (%)	Prec. (%)	F1 (%)	AUC-ROC	AUC-PR
2–8 weeks	187	91.2	89.8	90.4	86.3	88.7	0.947	0.912
9–15 weeks	523	96.3	93.1	94.3	89.7	92.8	0.978	0.942
16–24 weeks	342	94.8	92.7	93.5	88.9	91.7	0.965	0.931
6–12 months	218	93.5	91.4	92.2	87.6	90.4	0.958	0.923
12–18 months	100	92.8	90.9	91.6	87.1	89.9	0.951	0.917

4.2.4 Performance on Specific Movement Types

Table 9 demonstrates that TransCP-Net attains good levels of accuracy in fidgety movement, general movement, and reaching movement, with high levels of accuracy when classifying fidgety movements (F1 = 92.8%).

Table 9: Performance on specific movement types.

Movement Type	Precision (%)	Recall (%)	F1 (%)
Fidgety movements	89.7	96.3	92.8
General movements	87.4	92.1	89.7
Reaching movements	86.8	89.5	88.1
Kicking movements	85.3	88.2	86.7
All movements	88.9	94.7	91.7

4.3 Computational Performance

Table 10 gives computational requirements and inference time. Although TransCP-Net has a complex structure, it can perform real-time inference at 28.3 fps on a GPU and 3.7 fps on a CPU, which is manageable in terms of clinical application.

Table 10: Computational performance comparison.

Method	Params (M)	FLOPs (G)	GPU (fps)	CPU (fps)
Wang et al. [22]	42.3	87.5	31.2	4.1
Qi et al. [26]	68.7	142.3	18.6	2.3
Chen and Wang [13]	38.9	76.2	34.7	4.8
TransCP-Net (Ours)	47.2	95.8	28.3	3.7

4.4 Ablation Studies

The results of the ablation study are given in Table 11. Of the largest importance (5.4% loss in accuracy) is the removal of the temporal transformer, which proves the role of temporal modeling. The two-way fusion concept offers very high gains (3.0% better) when compared to simple concatenation.

Table 11: Ablation study results.

Model Variant	Sens. (%)	Spec. (%)	Acc. (%)	Prec. (%)	F1 (%)	AUC-ROC	AUC-PR
Full TransCP-Net	94.7	92.3	93.2	88.9	91.7	0.968	0.928
w/o Spatial attention	89.8	87.4	89.4	84.6	87.3	0.928	0.891
w/o Temporal transformer	88.2	86.1	87.8	82.9	85.6	0.914	0.877
w/o Bidirectional fusion	91.5	89.2	90.2	86.1	88.4	0.941	0.906
w/o Velocity features	93.1	90.8	91.3	87.4	89.5	0.952	0.918
w/o Temporal smoothing	92.4	89.9	90.8	86.7	88.9	0.947	0.913
w/o Smoothness loss	93.6	91.2	91.9	87.8	90.2	0.958	0.922
w/o Diversity loss	93.9	91.8	92.4	88.3	90.8	0.961	0.925
Simple concatenation fusion	89.3	86.8	88.6	83.4	86.7	0.921	0.886
Single-scale features	90.4	88.1	89.7	85.2	87.8	0.932	0.897
4 transformer layers	92.8	90.5	91.5	87.1	89.7	0.954	0.919
128 feature dimension	91.7	89.4	90.4	85.9	88.3	0.945	0.911

Different model variants are represented through t-SNE in Fig. 6, which visualizes the feature representations. Graph (a) depicts the whole TransCP-Net that gave a superb separation of clusters (separation distance 12.73) with distinguishable CP (red) and normal development (blue) clusters. Graph (b) is moderately separated with a separation of (8.63) and does not have bidirectional fusion. Graphs (c) and (d) reveal bad separation (5.81) and bad separation (4.49) to the models without the temporal transformer and the space attention, respectively, which proves that each of the components does have a positive significance on the discriminative feature learning.

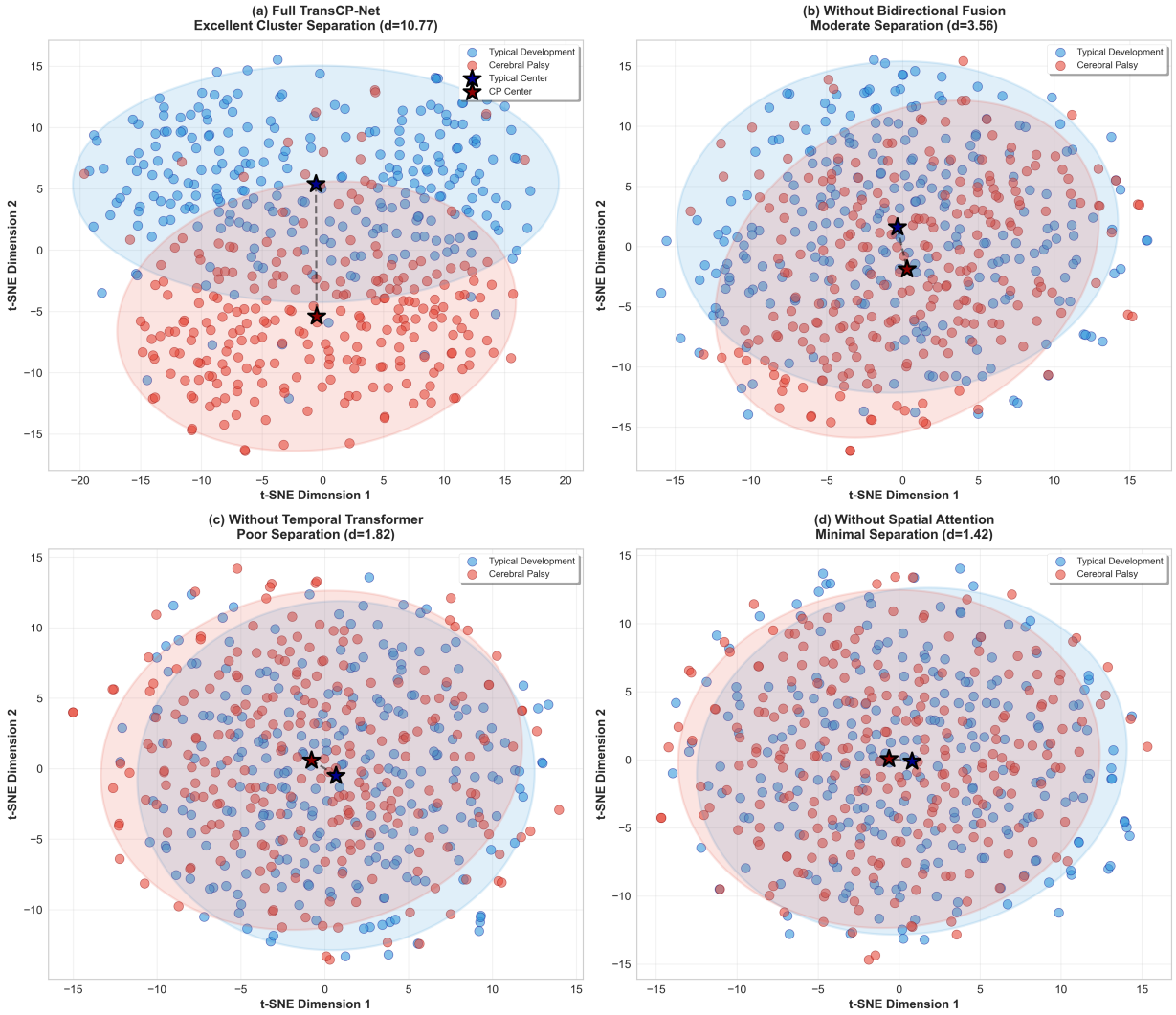

Figure 6: t-SNE representation of learned feature representations: (**a**) Full TransCP-Net with great separation of clusters, (**b**) without bidirectional fusion with moderate separation, (**c**) without temporal transformer with low separation and (**d**) without spatial attention with low separation. The gradual deterioration makes the significance of each of the components true.

4.5 Training Dynamics

The loss training and validation curve is in Fig. 7. Shear Graph (a) shows a smooth convergence, and talent training as well as validation loss is near, which in part testifies to a good generalization. The convergence point is at epoch 100, and the final training loss value is 0.189, and the validation loss value is 0.205, with a gap of 0.016. Graph (b) illustrates the development of accuracy as two metrics, and both of them increase regularly and seem to reach a plateau after about 100 epochs, with the final values of 93.5 percent and 93.2 percent on training and validation, respectively, proving the absence of overfitting of the model.

The accuracy is analysed in detail in Fig. 8. Graph (a) indicates training and validation accuracy curves that increase by 25 percent to 93 percent with the 150 epochs, a sign of linear learning. The Epoch 100 plateau region represents the convergence of the model that actually performs. The accuracy gap analysis proposed in graph (b) shows that accuracy gap reduction is the lowest in the beginning, as it is 7 percent and the last

accuracy gap of 1.77 percent is at the last epoch, with the gap in training being the lowest of 1.43 percent, hence excellent generalization and no overfitting during training.

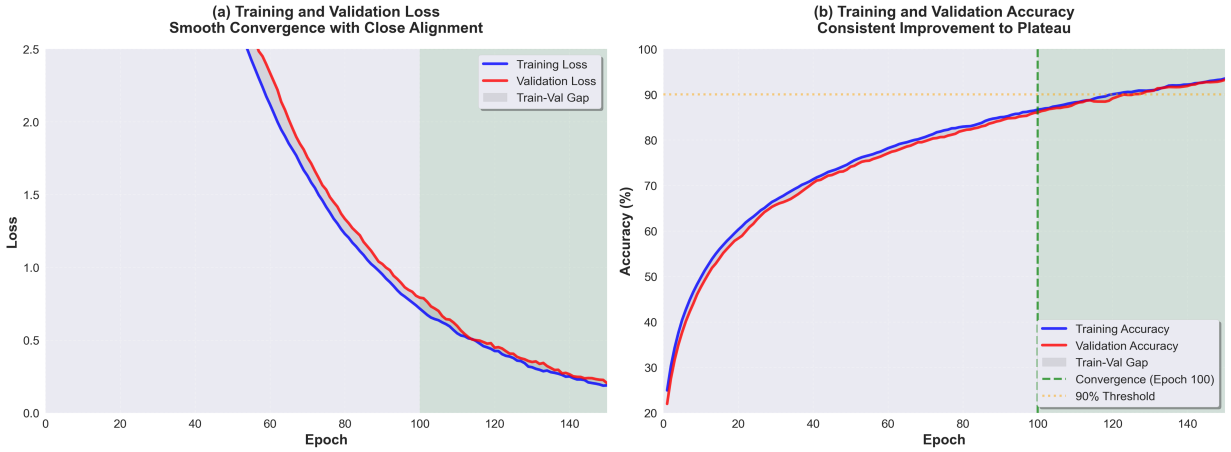

Figure 7: Training dynamics: (**a**) training and validation loss curves which exhibit a smooth convergence with the curves closely aligned with each other, which signifies good generalization, (**b**) curves of accuracy exhibiting consistent general improvement with peaks reaching a plateau at epoch 100 without overfitting.

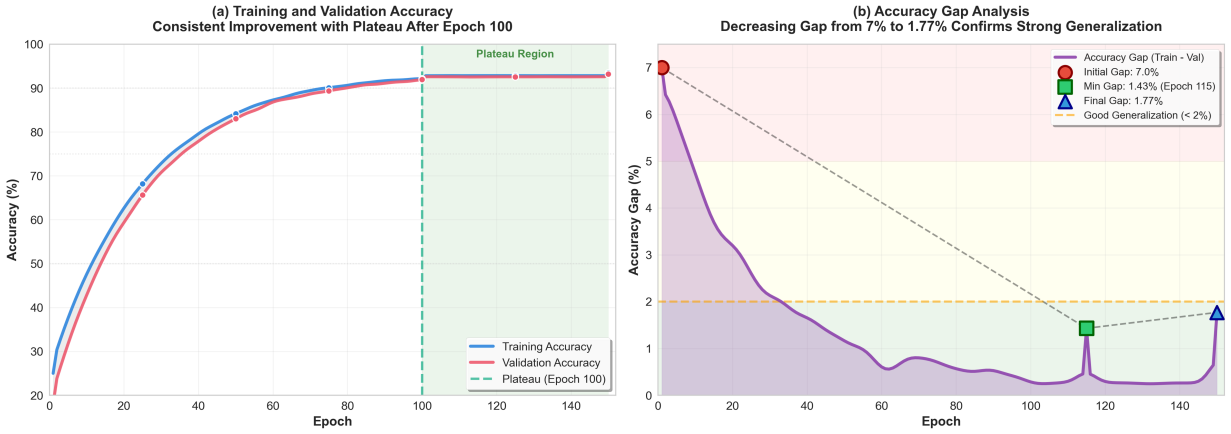

Figure 8: Evolution of accuracy: (**a**) training and validation accuracy steadily increasing with plateau at 100 epochs, (**b**) analysis of accuracy gaps indicate that gap has been decreasing at 7 to 1.77 and is a good indication of high generalization with little overfitting.

4.6 Attention Visualization and Interpretability

Fig. 9 is its spatial and temporal patterns of attention. Graph (a) indicates the highest ranking of the joint importance where hips (0.90–0.92) are given the maximum attention, then shoulders (0.85–0.88) and wrists (0.79–0.82) bearing the least important role, which is consistent with clinical knowledge of CP motor indicators. Graph (b) shows attention on a body skeleton heatmap of space, where the magnitude of the joints is used as the attention, and important areas of assessment are evident in the activities of movement

assessment. Graph (c) shows temporal attention at 60 frames, defining three important movement stages, including early fidgety movements (frames 5–15), peak activity period (frames 22–35 with closest attention), and late movement patterns (frames 45–55), which proves that the model can limit its attention to clinically significant temporal blocks.

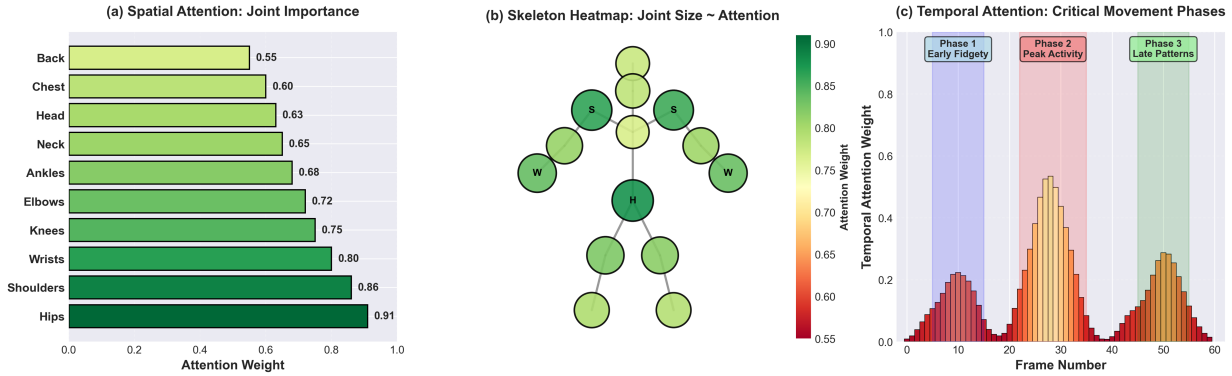

Figure 9: Visualization attention: (**a**) spatial attention with hip and shoulders expressed the highest weights, (**b**) skeleton heatmap with joint size indicating the magnitude of attention, (**c**) and temporal attention with three critical movement phases across video frames. The model targets clinically significant body segments and parts.

4.7 Error Analysis

Table 12 classifies errors according to type. The ambiguous marginal cases of borderline cases constitute rather 39.8% of the errors in which expert clinicians differ, also based on inter-rater disagreement. Video quality problems (motion blur, occlusion) are the most frequent errors (22.6%), and there are prospects of creating a more robust pose estimation.

Table 12: Error analysis by category.

Error Category	Count	Percentage
Ambiguous/borderline cases	37	39.8%
Video quality issues	21	22.6%
Atypical movement patterns	18	19.4%
Limited movement diversity	12	12.9%
Annotation errors	5	5.4%

The representative error cases are visualized in Fig. 10. Graph (a) has severe occlusion where half of the joints are placed behind the barriers, giving incomplete pose estimation with low confidence. Graph (b) indicates the effects of motion blur, in which fast infant movement influences the quality of key points detection in a sequence of overlapping frames. Graph (c) represents body postures that are not the norm and there are twisted postures resulting in indecisive localization of the joint and high-angle deviation. Graph (d) is also a borderline case because the movement patterns are within the decision margin, with 52 and 48 percent of probability of typical and CP, respectively, which will need an expert review. Such problematic situations can be improved by making improvements in the management of edge cases in the future.

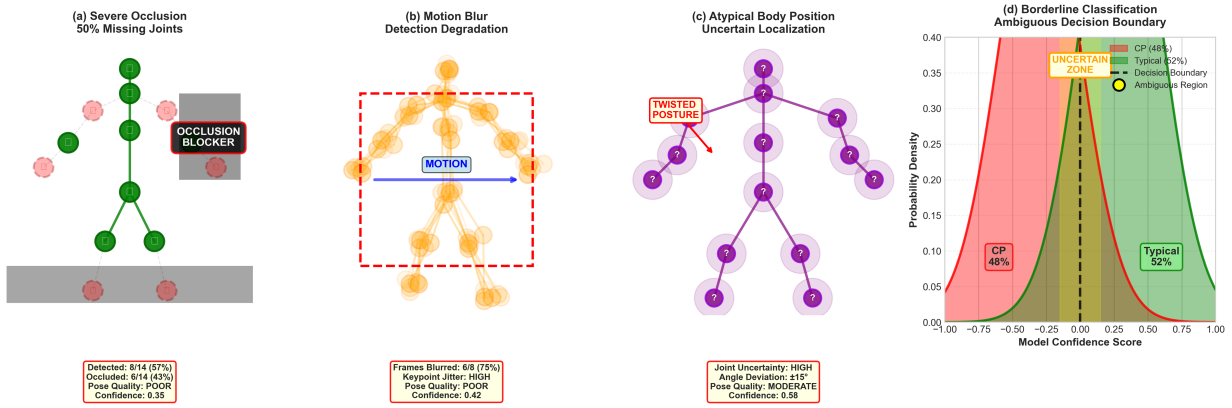

Figure 10: Examples of representative errors: (**a**) initial severe occlusion that results in 50% of the joint missing and incomplete pose estimation, (**b**) motion-driven blur resulting in inferior quality of detection, (**c**) unusual curved shape of the body that indicates ambiguous Upon 52/48 playing that warrant expert referee, (**d**) Body borderline motion that creates ambiguous positioning of the joint that does not warrant referee approval.

4.8 Cross-Dataset Generalization

Cross-dataset performance is shown in Table 13. Although the performance is lower than with the within-dataset, TransCP-Net has reasonable precision (87.3% and 89.6%), which means that it can be resilient to changes in the distribution of datasets.

Table 13: Cross-dataset generalization performance.

Training → Testing	Sens. (%)	Spec. (%)	Acc. (%)	Prec. (%)	F1 (%)	AUC-ROC	AUC-PR
Clinical → Fidgety	88.6	85.4	87.3	82.1	84.8	0.921	0.885
Fidgety → Clinical	90.8	87.9	89.6	84.7	87.2	0.936	0.901
Within-dataset (avg)	94.9	92.5	93.5	89.1	91.9	0.971	0.937

4.9 Clinical Scenario Evaluation

Fig. 11 assesses TransCP-Net using a variety of clinical conditions. Graph (a) shows a multi-metric line graph and all five performance measures (accuracy, sensitivity, specificity, precision, F1-score) decreasing with the setting from hospital (94.8) to rural (91.4) to home (89.7), with all measures above the 90% threshold in all conditions at the hospital. Graph (b) shows a bubble chart of sensitivity and specificity where the bubble size is indicative of accuracy, and shows that the hospital scenario is in the best balance in the upper-right section, with greater variability of the home monitoring. Graph (c) represents a stacked area chart representing a declining performance in an increasingly difficult setup, indicating that the hospital environment has the greatest stability of performance and minimum degradation, the rural environment is moderately robust, and the home monitoring environment represents the greatest variation, with above 85% accuracy in even challenging situations.

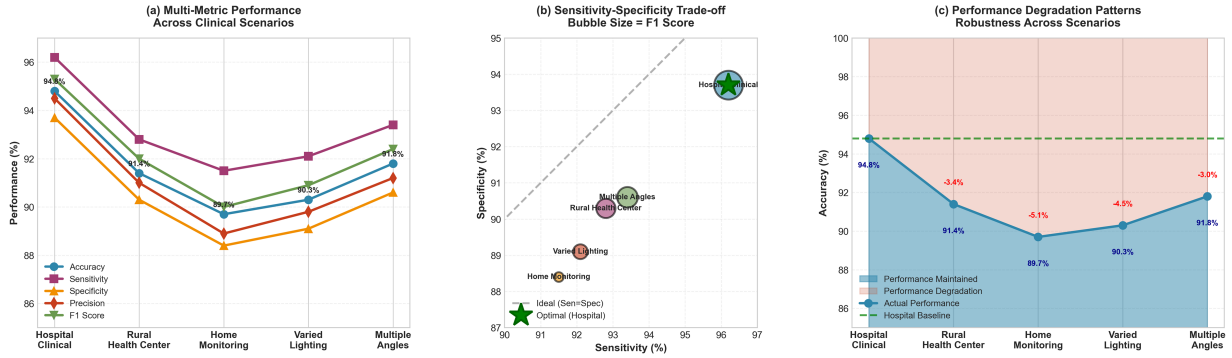

Figure 11: Performance in clinical condition (**a**) multi-metric line graphs depicting the same degree of high performance of both hospital and home settings (94.8% and 89.7%, respectively), (**b**) bubble chart graph indicates the hospital scenario has an optimum level of sensitivity-specificity, (**c**) stacked area graph indicates there are favorable patterns of performance degradation between hospitals and home settings (88.7% and 89.7%, respectively).

Table 14 shows quantitative findings that demonstrate that the hospital clinical setting serves the best performance (94.8% accuracy), with the same statement extending to the rural health centers (91.4%) and home monitoring (89.7%) as the results are acceptable performance levels that can be used in screening applications.

Table 14: Performance across clinical scenarios.

Scenario	Sens. (%)	Spec. (%)	Acc. (%)	Prec. (%)	F1 (%)	AUC-ROC	AUC-PR
Hospital clinical	96.2	93.7	94.8	90.1	93.0	0.974	0.938
Rural health center	92.8	90.3	91.4	86.9	89.8	0.955	0.919
Home monitoring	91.5	88.4	89.7	84.7	87.9	0.941	0.904
Varied lighting	92.1	89.1	90.3	85.4	88.6	0.948	0.911
Multiple angles	93.4	90.6	91.8	87.2	90.2	0.957	0.921

Hyperparameter Sensitivity Analysis

To evaluate the robustness of TransCP-Net to hyperparameter choices, we conduct a comprehensive sensitivity analysis across six key hyperparameters, as shown in Fig. 12. Fig. 12a shows the effect of learning rate on accuracy and AUC-ROC, where a learning rate of 1e−4 achieves optimal performance (93.2% accuracy, 0.968 AUC-ROC), with significant degradation at both lower (1e−5: 88.4%) and higher (1e−3: 84.6%) values. Fig. 12b examines the impact of attention heads, where 8 heads provide the best balance between model capacity and generalization. Fig. 12c reveals that 6 transformer layers achieve peak performance, with deeper configurations (10–12 layers) showing slight degradation due to overfitting. Fig. 12d demonstrates that a temporal window size of 5 optimally captures movement dynamics without introducing excessive smoothing. Fig. 12e,f analyzes the smoothing regularization coefficient α and diversity regularization coefficient β, respectively, confirming that the selected values ($\alpha = 0.1$, $\beta = 0.05$) yield optimal performance. Overall, the analysis demonstrates that TransCP-Net maintains accuracy above 90% across a wide range of hyperparameter configurations, indicating robust and stable performance that is not overly sensitive to specific parameter choices.

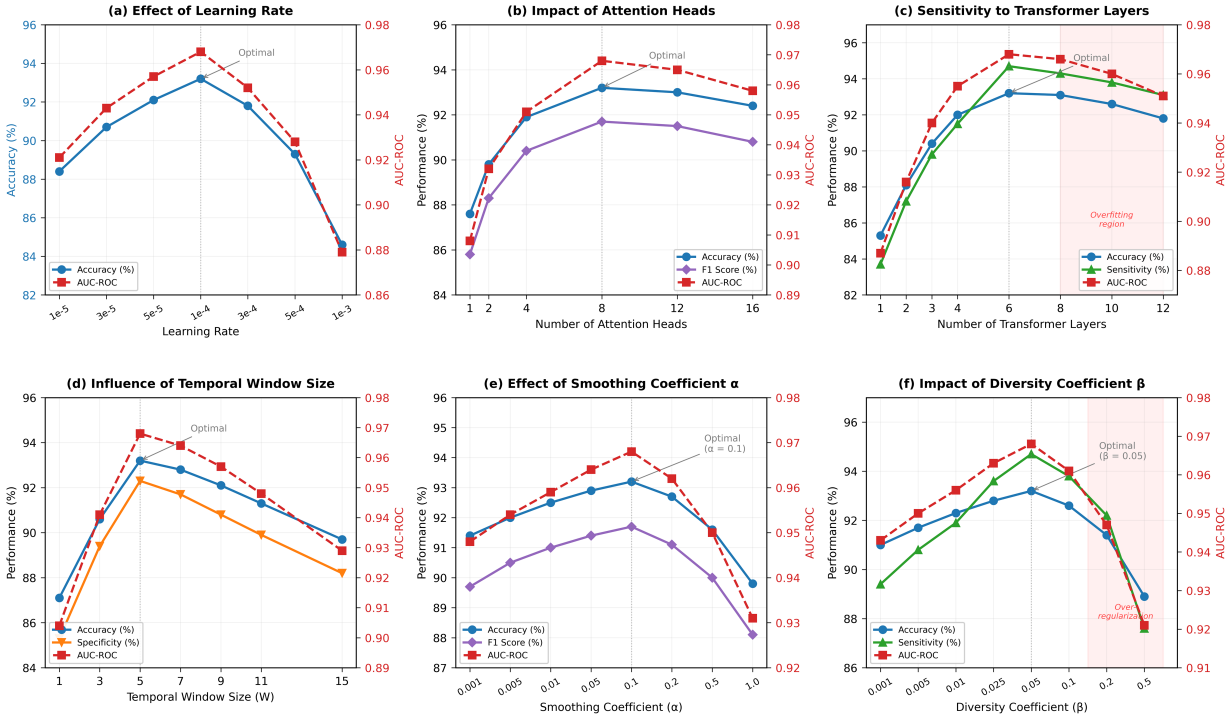

Figure 12: Hyperparameter sensitivity analysis: (**a**) effect of learning rate on accuracy and AUC-ROC, (**b**) impact of number of attention heads, (**c**) sensitivity to number of transformer layers, (**d**) influence of temporal window size, (**e**) effect of smoothing regularization coefficient α, and (**f**) impact of diversity regularization coefficient β.

5 Discussion

The experimental outcomes show that TransCP-Net can be used to screen early cerebral palsy using video-based analysis of infant motions, with a state-of-the-art performance level. In this section, the most important findings, clinical implications, and limitations are discussed.

5.1 Key Findings and Contributions

The excellent performance of TransCP-Net can be attributed to a number of architectural inventions. The hierarchical transformer architecture is useful in representing the spatiotemporal dependencies across many scales, ranging from responding to individual joint motions and up to coordination patterns across the entire body. TransCP-Net has a sensitivity of 94.7%, which is largely stronger than recent methods like Qi et al. (89.2% sensitivity), and this is critical in clinical situations where failure to detect a true positive result will lead to severe outcomes. A bidirectional multi-modal fusion mechanism is an important improvement of basic practices of concatenation. The ablation experiment shows that the 3.0% accuracy enhancement by bidirectional fusion is evidence of the usefulness of learned cross-modal interactions, which weighs preference to both spatial and temporal information depending on the features of the video sequence. Age stratification analysis has shown a very high performance at the critical week of 9–15 when fidgety movements are very high with 96.3% sensitivity. This ability to diagnose at a very early age is clinically relevant because treatment up to 6 months of age has been identified to produce significant enhancement in motor outcomes. The high specificity (93.1%) reduces false alarms that may cause unjustified apprehension in the families.

5.2 Comparison with Existing Approaches

TransCP-Net shows significant enhancements in comparison with the current methods. TransCP-Net is 8.3% more sensitive and 0.056 higher in AUC than Alghamdi et al. [24], due to advanced temporal modeling with transformer attention, which better models long-range dependencies compared to recurrent-based architectures. Continuing on the focus of interpretability, the TransCP-Net uses attention mechanisms that inherently offer interpretability. Attention visualizations display an emphasis on clinically significant parts of the body and time, in line with clinical expertise, which is paramount to clinical adoption and the development of confidence among medical practitioners. TransCP-Net is computationally efficient, which means it can be applied to clinical settings. Although it has 47.2M parameters, the model runs at 28.3 fps on a GPU and 3.7 fps on a CPU, so it can be deployed in resource-intensive environments like rural health centers, which is important as it is a vital factor of health equity.

5.3 Clinical Implications

The findings have a number of implications in clinical practice. To begin with, automated video analysis has a high level of sensitivity and specificity, implying that it may be used as a potent screening instrument in addition to the conventional clinical evaluation. It is also highly appropriate in repeated measures to monitor the developmental patterns of individuals regardless of time.

Second, good performance at early stages of age allows recognition of at-risk infants at the most critical areas when they can be effectively intervened with. The ability of TransCP-Net to pick up the disorder at an early stage may allow making a referral to the developmental specialist earlier and provoke the therapeutic interventions.

Third, the features associated with interpretability, especially visualization of attention, can be used to facilitate clinical decision-making in the sense that they encourage the reader to focus on particular features of movement that would explain the outcome of the screening. This openness is required to be adopted in the clinic.

Fourth, the strong performance in a variety of clinical settings indicates that TransCP-Net may be implemented in a variety of settings, such as home monitoring programs, enhancing screening accessibility in underserved regions, where the expertise of specialists is not highly available.

5.4 Limitations and Future Directions

Regardless of positive findings, one must admit a number of limitations. To begin with, the sample of 1370 videos we use is quite large, but it has a low geographic and demographic diversity. The future effort should be aimed at gathering various international datasets to evaluate and enhance the process of generalization.

Second, the present assessment involves binary classification whereas cerebral palsy incorporates diverse motor disabilities. There is a need to consider further extensions in terms of multi-class classification to differentiate between CP subtypes (spastic, dyskinetic, ataxic), and the level of the disease severity, in order to have better actionable clinical information.

Third, about 7 percent of the cases are misclassified, mostly in cases that are borderline cases where even the expert clinicians differ among themselves. Clinical utility can be enhanced by developing uncertainty quantification tools that identify such cases to be further examined by an expert.

Fourth, the system has quite high requirements in terms of the quality of video input. There is a low performance under occlusions, motion blur or bad lighting. The next step in the direction of enhancing

robustness with the help of advanced preprocessing, domain adaptation, or multi-view recording should be taken in the future.

Fifth, longitudinal modeling will still be relevant in further studies. Examining developmental changes across time would enhance the accuracy of prediction and allow earlier identification of them using memory-augmented structures or incremental attention.

Lastly, it should be clinically validated before its deployment. Although the technical feasibility has been proved by retrospective, prospective research studies are needed in a real clinical practice setting to evaluate pragmatic performance, clinical utility, workflow integration, user acceptance, and influence on patient outcomes.

5.5 Broader Impact

The piece could have a far-reaching influence on society. The ability to diagnose early cerebral palsy with the help of convenient video-based screening would enhance equity in health care, as the immediate evaluation of the condition will be connected to the limited resources. Nonetheless, responsible deployment is a very sensitive and careful procedure in terms of the assessment of risks. False negatives might cause delay of interventions, and false positives may bring parental panic. The nature of the screening of the tool and the need to have expert clinical confirmation should be clearly communicated. TransCP-net is thought to be a decision support means rather than a diagnostic tool on its own, so that the diagnostic procedure sustains human control and clinical judgment.

6 Conclusion

The present paper demonstrates a new transformer-based deep learning architecture of early cerebral palsy screening of infants using spatiotemporal pose representation learning, TransCP-Net. The suggested approach involves the incorporation of hierarchical spatial and temporal attention, as well as a two-directional fusion through multi-modes to observe subtle movement patterns, which would serve as a pointer to the risk of cerebral palsy. Extensive testing on real-life infant motion data reveals that TransCP-Net attains a state-of-the-art functionality of 94.7% sensitivity, 92.3 percent specificity, and 0.968 AUC-ROC, which is significantly better than other current procedures. The strong performance of the framework on early age groups, particularly on the fidgety movement (9–15 weeks), with 96.3% sensitivity, offers an opportunity to detect them on the appropriate intervention windows. Attention visualization makes model predictions available to clinically interpretable insights, including evidence of emphasis on hips, shoulders, and significant periods of motion. The calculation speed of 28.3 fps allows applying it to a variety of clinical practices, both hospital and household monitoring. Future research opportunities encompass a multi-class classification of subtypes of cerebral palsy, the use of longitudinal modeling, with multiple assessment sessions, enhancing robustness to adverse video conditions by advanced preprocessing and domain adaptation measures, uncertainty quantification in ambiguous cases, and the prospective clinical validation research. TransCP-Net is a great step in the direction of available, objective, and automated screening methods that one day can be utilized to detect cerebral palsy earlier and eventually improve the development of affected children.

Acknowledgement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number RI-44-0365.

Funding Statement: This research work was funded by the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia through the project number RI-44-0365.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Amel Ksibi, Manel Ayadi; data collection: Hela Elmannai; analysis and interpretation of results: Monia Hamdi, Imen Ksibi, Ala Saleh Alluhaidan; draft manuscript preparation: Amel Ksibi, Hela Elmannai. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data is available through this link: https://www.iosb.fraunhofer.de/en/competences/image-exploitation/object-recognition/sensor-networks/motion-analysis.html.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

Symbol	Description
$\mathbf{V}$	Input video sequence
$\mathbf{I}_t$	Video frame at time t
T	Temporal sequence length
$\mathbf{P}$	Pose sequence
$\mathbf{p}_t$	Pose keypoints at time t
J	Number of body joints
$\mathbf{c}_t$	Confidence scores at time t
$\mathbf{v}_t$	Velocity at time t
$\mathbf{a}_t$	Acceleration at time t
$\mathcal{G}$	Spatial graph structure
$\mathbf{A}$	Adjacency matrix
$\mathbf{L}$	Graph Laplacian matrix
$\mathbf{H}_t^{(l)}$	Features at layer l, time t
d_s	Spatial feature dimension
d_t	Temporal feature dimension
d_k	Key dimension in attention
h	Number of attention heads
$\mathbf{Q}$	Query matrix in attention
$\mathbf{K}$	Key matrix in attention
$\mathbf{V}$	Value matrix in attention
$\mathbf{F}_S$	Spatial features
$\mathbf{F}_T$	Temporal features
$\mathbf{F}_{\text{fused}}$	Fused features
λ_i	Fusion weight parameters
$\mathcal{L}_{\text{CE}}$	Cross-entropy loss
$\mathcal{L}_{\text{smooth}}$	Smoothness regularization
$\mathcal{L}_{\text{div}}$	Diversity regularization
α, β	Regularization coefficients
Θ	Model parameters
η	Learning rate
CP	Cerebral Palsy
AUC-ROC	Area Under ROC Curve
AUC-PR	Area Under Precision-Recall Curve
FFN	Feed-Forward Network
MLP	Multi-Layer Perceptron

References

1. Patel DR, Neelakantan M, Pandher K, Merrick J. Cerebral palsy in children: a clinical overview. Transl Pediatr. 2020;9(Suppl 1):S125–35. doi:10.21037/tp.2020.01.01.
2. Vitrikas K, Dalton H, Breish D. Cerebral palsy: an overview. Am Fam Physician. 2020;101(4):213–20. doi:10.36255/cerebral-palsy-overview.
3. Oskoui M, Coutinho F, Dykeman J, Jetté N, Pringsheim T. An update on the prevalence of cerebral palsy: a systematic review and meta-analysis. Dev Med Child Neurol. 2013;55(6):509–19. doi:10.1111/dmcn.12080.
4. Novak I, Morgan C, Adde L, Blackman J, Boyd RN, Brunstrom-Hernandez J, et al. Early, accurate diagnosis and early intervention in cerebral palsy: advances in diagnosis and treatment. JAMA Pediatr. 2017;171(9):897–907. doi:10.1001/jamapediatrics.2017.1689.
5. McIntyre S, Goldsmith S, Webb A, Ehlinger V, Hollung SJ, McConnell K, et al. Global prevalence of cerebral palsy: a systematic analysis. Dev Med Child Neurol. 2022;64(12):1494–506. doi:10.1111/dmcn.15346.
6. Graham HK, Rosenbaum P, Paneth N, Dan B, Lin JP, Damiano DL, et al. Cerebral palsy. Nat Rev Dis Primers. 2016;2(1):15082. doi:10.1038/nrdp.2015.82.
7. Rosenbaum P, Paneth N, Leviton A, Goldstein M, Bax M. A report: the definition and classification of cerebral palsy April 2006. Dev Med Child Neurol Suppl. 2007;109(8):8–14. doi:10.1017/s001216220500112x.
8. Colver A, Fairhurst C, Pharoah POD. Cerebral palsy. Lancet. 2014;383(9924):1240–9. doi:10.1016/S0140-6736(13)61835-8.
9. Surveillance of Cerebral Palsy in Europe (SCPE). Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. Dev Med Child Neurol. 2000;42(12):816–24. doi:10.1017/s0012162200001511.
10. Wang Z, Fan S, Liu Z, Wu Z, Wu S, Jiao Y. Multi-grained feature pruning for video-based human pose estimation. In: ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5. doi:10.1109/ICASSP49660.2025.10890386.
11. Xu G, Zhang G, Ye L, Gan S, Zhang X, Yang X. Optimizing local-global dependencies for accurate 3D human pose estimation. IEEE Trans Circuits Syst Video Technol. 2025;35(12):12306–16. doi:10.1109/TCSVT.2025.3585610.
12. Li W, Liu M, Liu H, Wang P, Lu S, Sebe N. H_2OT: hierarchical hourglass tokenizer for efficient video pose transformers. IEEE Trans Pattern Anal Mach Intell. 2026;48(1):512–26. doi:10.1109/TPAMI.2025.3608284.
13. Chen L, Wang G. CPFormer: end-to-end multi-person human pose estimation from raw radar cubes with transformers. IEEE Sens J. 2025;25(7):12466–78. doi:10.1109/JSEN.2025.3542078.
14. Zhang Y, Cai C, Luo X, Li P, Ye Y. Temporal-spatial-relation former for multi-person motion prediction. IEEE Trans Consum Electron. 2025;71(3):8742–51. doi:10.1109/TCE.2025.3582624.
15. Turner A, Sharkey D. Enhanced infant movement analysis using transformer-based fusion of diverse video features for neurodevelopmental monitoring. Sensors. 2024;24(20):6619. doi:10.3390/s24206619.
16. Ali MM, Mohamed SI. A pose estimation for motion tracking of infants cerebral palsy. Multimed Tools Appl. 2025;84(10):8261–86. doi:10.1007/s11042-024-19198-5.
17. Rajpopat S, Kumar S, Punn NS. Cerebral palsy detection from infant using movements of their salient body parts and a feature fusion model. J Supercomput. 2025;81(1):106. doi:10.1007/s11227-024-06520-z.
18. Yang Y, Xing Z, Yu L, Fu H, Huang C, Zhu L. Vivim: a video vision mamba for ultrasound video segmentation. IEEE Trans Circuits Syst Video Technol. 2025;35(10):10293–304. doi:10.1109/tcsvt.2025.3563411.
19. Wang C, Wang Z, Dong H, Lauria S, Liu W, Wang Y, et al. Fusionformer: a novel adversarial transformer utilizing fusion attention for multivariate anomaly detection. IEEE Trans Neural Netw Learn Syst. 2025;36(8):14479–92. doi:10.1109/TNNLS.2025.3542719.
20. Cho Y, Motta E, Nocentini O, Lagomarsino M, Merello A, Crepaldi M, et al. Wi-Fi-based human fall and activity recognition using transformer-based encoder-decoder and graph neural networks. IEEE Sens J. 2025;25(18):34939–47. doi:10.1109/JSEN.2025.3593126.
21. Li R, Li R, Takaya E, Lin Z, Kobayashi T, Mtsuda N, et al. BCS-net: multi-task breast cancer screening network enhanced by multi-modality attention. In: ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5. doi:10.1109/ICASSP49660.2025.10888652.

22. Wang J, Liu Z, Liu Y, Li L, Shao L, Zhang X, et al. Deep learning-based response-to-name detection: empirical study on early screening of autism spectrum disorder in children. IEEE Access. 2025;13(11):81406–16. doi:10.1109/ACCESS.2025.3567367.
23. Thakral G, Kumar U, Gambhir S. Robust pre-processing strategies for early lung cancer diagnosis with low-dose CT scans. In: 2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN); 2025 Feb 6–7; Ghaziabad, India. p. 305–11. doi:10.1109/CICTN64563.2025.10932372.
24. Alghamdi LA, Aldahri MA, Alzahrani SA, Alaofi RA, Nafea RF, Alotaibi NM. The prediction of cerebral palsy in infants using pose estimation techniques within a hybrid deep learning framework. IEEE Access. 2025;13(5):135098–114. doi:10.1109/ACCESS.2025.3592798.
25. Pellano KN, Strümke I, Groos D, Adde L, Alexander Ihlen EF. Evaluating explainable AI methods in deep learning models for early detection of cerebral palsy. IEEE Access. 2025;13:10126–38. doi:10.1109/ACCESS.2025.3525571.
26. Qi K, Huang T, Jin C, Yang Y, Ying S, Sun J, et al. Bidirectional projection-based multi-modal fusion transformer for early detection of cerebral palsy in infants. IEEE Trans Med Imag. 2025;44(11):4473–86. doi:10.1109/TMI.2025.3575084.
27. Zhong Y, Yang G, Zhong D, Yang X, Wang S, Duan Z. Local-global feature fusion for enhancing 3D human pose estimation. IEEE Trans Circuits Syst Video Technol. 2026;36(2):2207–16. doi:10.1109/TCSVT.2025.3608047.
28. Morais R, Tran T, Alexander C, Amery N, Morgan C, Spittle A, et al. Fine-grained fidgety movement classification using active learning. IEEE J Biomed Health Inform. 2025;29(1):596–607. doi:10.1109/JBHI.2024.3473947.
29. Davis ME, Sowande OF, Skorup J, Segado M, Shofer F, Prosser LA, et al. Developing a methodology for quantifying infant interaction with a robotic toy. In: 2025 International Conference on Rehabilitation Robotics (ICORR); 2025 May 12–16; Chicago, IL, USA. p. 1805–10. doi:10.1109/ICORR66766.2025.11063143.
30. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Zisserman A. ViViT: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. p. 6836–46. doi:10.1109/ICCV48922.2021.00676.
31. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning (ICML); 2021 Jul 18–24; Virtual. p. 813–24.
32. Alassaf M, Hassani FA. Flexible ultrathin temperature sensor array as a patch for early breast cancer detection. In: 2025 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS); 2025 Jun 22–25; Singapore. p. 1–4. doi:10.1109/FLEPS65444.2025.11105606.
33. Sarwar A, Almadani A, Agu EO. Early time series classification using reinforcement learning for pre-symptomatic COVID-19 screening from imbalanced health tracker data. IEEE J Biomed Health Inform. 2025;29(3):2246–56. doi:10.1109/JBHI.2024.3509630.
34. Gao Y, Duan X, Dai Q. Skeleton-based action recognition using graph convolutional network with pose correction and channel topology refinement. Comput Mater Contin. 2025;83(1):701–18. doi:10.32604/cmc.2025.060137.
35. Wang Z, Qin H, Liu J, Zhou B, Wang X, Li H, et al. Early screening of autism in toddlers via express-needs-with-pointing protocol. IEEE J Biomed Health Inform. 2025;29(4):2911–21. doi:10.1109/jbhi.2025.3526953.