A Lightweight Deep Residual Network for Rehabilitation Activity Recognition in Heterogeneous Pediatric Populations

Simone Costantini [1 2] **Benedetta Giachetti** [3 4] **Fabio Alexander Storm** [2] **Emilia Biffi** [2] **Elena Mugellini** [3] **Anna Maria Bianchi** [1] **Giuseppe Andreoni** [2 5]

Abstract

Human Activity Recognition based on wearable Inertial Measurement Units (IMUs) has emerged as a promising technology for the automated quantification of rehabilitation dosage and rehabilitation activity identification. However, existing solutions rely on multi-sensor configurations limiting clinical usability or fail to generalize to populations with heterogeneous motor functions. This study developed a lightweight residual network with self-attention mechanisms for classifying different phases of the rehabilitation activities (Rest, Balance, Walk) using data from a single IMU placed at the lower back, and collected from a pediatric cohort including 10 neurotypical children (mean age: 8.7 ± 2.5 years, 5 females) and 8 patients with neuromotor disorders as a consequence of cerebral palsy or acquired brain injury (mean age: 10.4 ± 3.2 years, 4 females). A preliminary ablation study across different IMU channel combinations revealed that combining accelerometer, gyroscope, and magnetometer signals allowed the model to achieve the best performance, with the magnetometer providing a key contribution for better discriminating between low-dynamic activities (Rest and Balance). Based on the optimal channel configuration identified in the ablation study, a Leave-One-Subject-Out cross-validation framework proved the model generalization abilities across heterogeneous motor functional domains, achieving an average macro F1-score of 0.81. These results confirm that the proposed framework provides an ecological and reliable tool for the objective recognition and quantification of rehabilitation activity in a clinical context.

[1]Department of Electronics Information and Bioengineering, Politecnico di Milano, Milan, Italy [2]Scientific Institute, IRCCS "E.Medea", Bosisio Parini, Italy [3]HumanTech Institute, Haute école spécialisée de Suisse occidentale, 1700, Fribourg, Switzerland [4]Department of Computer Science, University of Fribourg, Fribourg, Switzerland [5]Department of Design, Politecnico di Milano, Milan, Italy. Correspondence to: Simone Costantini <simone.costantini@polimi.it>.

Proceedings of the Swiss AI Days 2026, Martigny, VS & Fribourg, FR, Switzerland PMLR 309, 2026.

1. Introduction

Pediatric neuromotor disorders are among the most prevalent causes of lifelong physical disability, which have a huge impact on the subject's motor functions and independence (Steinmetz et al., 2024). In this context, motor rehabilitation is crucial for motor function recovery (Cao et al., 2014). According to evidence-based guidelines, high-intensity and goal-oriented training is essential to promote neuroplasticity and improve motor learning (Novak et al., 2020), and can be supported by robot-based devices. Also, in relation to the optimal challenge point framework, motor learning is more effective when the rehabilitation task difficulty is tailored to the patient's motor functional capacity (Guadagnoli & Lee, 2004). As a consequence, measuring both the quantity and quality of motor activity becomes critical to guarantee that the rehabilitation dosage is adequate to induce functional recovery and to meet the patient's residual motor capabilities.

However, the objective quantification of the rehabilitation dosage is still an open challenge. Specifically, current clinical practice relies largely on subjective observational scales (González Barral & Servais, 2025), which are time-consuming and are not particularly suited to capture the granular details of rehabilitation intensity, such as the actual time spent by the patient actively doing motor exercises compared to passive rest phases (Cope & Mohn-Johnsen, 2017). The lack of a quantitative feedback limits the therapists' ability to personalize therapies effectively and to monitor patient adherence to the treatment plan (Fundarò et al., 2018).

To replace the need for manual observation and facilitate the automated identification of motor activities, Human Activity Recognition (HAR) based on Wearable Inertial Measurement Units (IMUs) has recently emerged as a promising technological solution for ecological motor activity assessment in rehabilitation contexts too (Kańtoch, 2017). While

preliminary studies relied on traditional machine learning approaches using handcrafted features to classify human activities (Ahmadi et al., 2020), Deep Learning (DL) architectures have quickly revolutionized this research domain (Chen et al., 2021). Specifically, DL models based on convolutional and recurrent neural networks are able to learn complex feature representations directly from raw IMU data, reaching a consistent improvement in classification performance across many HAR tasks (Kim et al., 2021; Oleh et al., 2024). More recently, sophisticated architectures such as residual networks (ResNets) and attention-based models have proven the highest performance, being particularly effective in capturing long-term temporal dependencies and spatial correlations within IMU sensors, thus unlocking complex activity recognition tasks (Mekruksavanich et al., 2022; 2025).

Nevertheless, despite these technological advances, the deployment of DL-based HAR systems into pediatric clinical practice still encounters critical issues. The main barrier lies within the domain shift problem since DL models trained on datasets comprising data from adult or neurotypical underage subjects often fail to generalize to pediatric patients, where inter-subject motor variability is higher by nature (Khaked et al., 2025; Tørring et al., 2024). Additionally, many high-performing HAR solutions require complex multi-sensor setups that are hard to use with children in a clinical context (Oleh et al., 2024), reducing the ecological validity of the rehabilitation session and, as a result, the deterioration of the patients' compliance and adherence to the treatment plan.

Therefore, to face these challenges, the current study aims to develop and validate a lightweight deep residual network for automated classification of rehabilitation activities (i.e., Rest, Balance, Walk) in a heterogeneous pediatric population including both neurotypical subjects and patients with neuromotor disorders. The proposed solution will focus on reaching high generalization capabilities across subjects with different motor functional domains using IMU data from one single device, thus providing an ecological tool for the objective quantification of rehabilitation exercises dosage in a clinical environment.

2. Materials and Methods

2.1. Participants

The present study involved a total of 18 participants, divided into two groups: 10 neurotypical underage subjects (mean age: 8.7 ± 2.5, 5 females), and 8 pediatric patients (mean age: 10.4 ± 3.2, 4 females) with neuromotor disorders as a consequence of cerebral palsy or acquired brain injury (for demographic and clinical-related information, see Appendix A). The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee Lombardia 2 (protocol code: L2-246; date of approval: March 22nd, 2025). The study is registered at ClinicalTrials.gov (identifier: NCT06993389). Participants' guardians signed a written informed consent. All data were pseudonymized.

2.2. Data acquisition

The eight underage participants with neuromotor disorders performed 20 rehabilitation sessions, according to their individual treatment plan. However, data acquisition was limited to two or three sessions spread across the treatment period per patient to avoid reducing therapy acceptance. Conversely, neurotypical participants went through a single session.

All sessions were conducted at the Scientific Institute, IRCCS "E. Medea" using the Gait Real-time Analysis Interactive Lab (GRAIL, Motek Medical, The Netherlands) system, a semi-immersive augmented reality medical device equipped with a two degrees of freedom motion frame, integrated force plates, a motion-capture system composed of 10 optoelectronic cameras, a dual-belt treadmill, and a 180° cylindrical projection screen.

Sessions included four to six lower-limb rehabilitation exercises, each lasting four to ten minutes, typically divided equally between balance exercises (hereafter, Balance), and gait exercises (hereafter, Walk). The specific exercises were chosen by the therapist for each participant and session: for patients, the selection was driven by their treatment goals, while for neurotypical subjects, exercises were suited to their motor abilities to provide an appropriate level of physical challenge. Additionally, before starting the first exercise of the session, participants were asked to relax and remain calm for approximately five to ten minutes (hereafter, Rest). This period was used both for the participant's initial assessment and for the technical setup of the GRAIL system by the therapist. The start and end timestamps of each activity (i.e., Rest, Balance, Walk) were manually annotated by the research team to facilitate the following labeling process.

Throughout the rehabilitation session, nine-channel IMU data (i.e., three-axis accelerometer, three-axis gyroscope and three-axis magnetometer) was collected at 128 Hz through the Shimmer3 IMU sensor unit (Shimmer Sensing, Dublin, Ireland), which was placed on the participants' lower back (L5 vertebra) with an elastic band.

2.3. Data preparation

IMU data from each participant and rehabilitation session was conditioned channel-wise by means of a second-order infinite impulse response Butterworth band-pass filter with low cut-off frequency set at 0.05 Hz to remove sensor

drift and high cut-off frequency at 10 Hz to attenuate high-frequency noise. Then, the nine-channel IMU signals were segmented into 3-second sliding windows with 67% overlap, such that each window represented an individual sample for the three-class classification task.

Labeling was done following a majority-rule approach: each window was assigned the label of a target activity (i.e., Rest, Balance, or Walk) if at least 90% of the window time points corresponded to one activity. Windows including mixed activities that did not meet this criterion were excluded to improve dataset quality and reduce ambiguity. The final dataset consisted of about 27% samples belonging to Rest, 29% to Balance, and 44% to Walk.

Finally, each window was z-score normalized, with mean and standard deviation parameters computed exclusively from the training set data.

2.4. Model architecture and training parameters

The rehabilitation activity classification task was performed by a custom lightweight ResNet architecture (Figure 1) having a total of about 300 thousand trainable parameters. The model is composed of an encoder for high-level feature extraction, a self-attention pooling layer to synthesize temporal information, and a classification head for final classification into the three target activities.

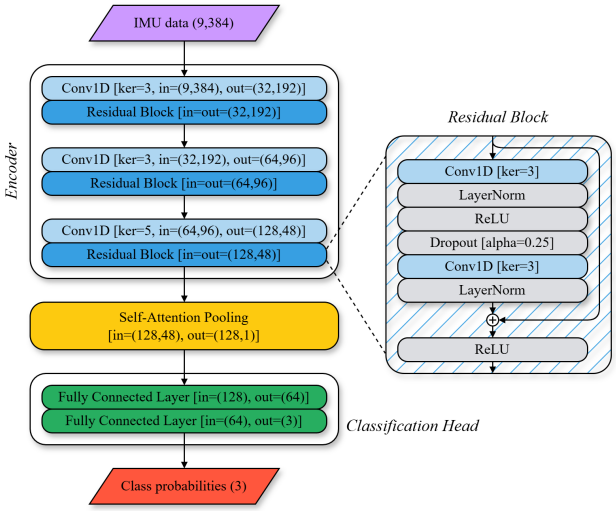

Figure 1. The architecture of the proposed deep residual network model for rehabilitation activity classification.

The encoder consists of three stages, each of them including a one-dimensional convolutional layer and a residual block. The convolutional layer has kernel size set to 3 for the first two stages and to 5 for the third stage, and the stride parameter set to 2 for progressive temporal downsampling. Conversely, the residual block consists of two convolutional layers with ReLU activations, dropout regularization ($\alpha = 0.25$), and layer normalization. Following the original architecture proposed in (He et al., 2016), residual connections allow for identity mapping between input and output, facilitating the learning of deep high-level features without gradient degradation.

Downstream of the encoder, a self-attention pooling layer is used to synthesize information along the temporal dimension. Specifically, the 128-dimensional embeddings at each downsampled timestamp attend to one another, thus enabling the learning of long-range temporal dependencies within the 3-second window (Vaswani et al., 2017). The resulting attended embeddings are weighted-averaged across the temporal dimension to produce a 128-dimensional feature vector.

The self-attention pooling layer output is finally sent to the classification head, which consists of two fully connected layers: the first layer compresses the 128-dimensional representation into a 64-dimensional latent space using a ReLU activation and dropout ($p = 0.25$); the second layer maps these latent features to the three target rehabilitation activities through a Softmax activation, producing the probability distribution for each class.

The model was trained using the Adam optimizer, with a learning rate of 10^{-6}, a weight decay of 10^{-5} for regularization, and a batch size of 128. A weighted cross-entropy loss, with weights computed as the inverse of class frequencies in the training set, was applied to handle non-negligible class imbalance. Also, to further address overfitting, the training loop implemented early stopping with patience of 20 epochs based on the macro F1-score on the validation set, and an adaptive learning rate scheduler with patience of 8 epochs to halve the learning rate when F1-macro plateaued.

Model training was run on a PC with an Intel Core i7-10750H CPU @ 2.6 GHz and an NVIDIA GeForce GTX 1650 Ti GPU. The entire pipeline was implemented in Python 3.10. Specifically, the `SciPy` library was exploited for data preparation, `PyTorch` was used for model architecture definition and training, and `scikit-learn` was employed to compute classification performance metrics.

2.5. Ablation study

To determine the optimal IMU channel configuration for rehabilitation activity recognition, an ablation study was performed across all seven combinations of the three IMU sensors types: accelerometer only (*Acc*), gyroscope only (*Gyro*), magnetometer only (*Mag*), accelerometer and gyroscope (*Acc+Gyro*), accelerometer and magnetometer (*Acc+Mag*), gyroscope and magnetometer (*Gyro+Mag*), and all sensors enabled (*Acc+Gyro+Mag*).

A group 6-fold cross-validation was used to ensure that data belonging to the same participant entirely fell exclusively

into the training, validation or test sets. In each fold, a partition of 3 participants was held out as a test set, while the remaining 15 participants formed the development set. This set was further split such that 12 participants were used for training and 3 for internal validation, maintaining a subject-independent approach throughout the ablation study.

The best-performing sensor configuration was chosen by ranking the seven combinations based on their median macro F1-score. Statistical significance of differences across configurations was assessed using a Friedman test followed by Durbin-Conover post-hoc pairwise comparisons ($p < 0.05$). All statistical analyses for the ablation study were performed using R (R 4.4.1, Vienna, Austria).

2.6. Model validation and deployment

The best-performing sensor configuration was selected for the validation study. A Leave-One-Subject-Out (LOSO) cross-validation approach was used to assess the ability of the model to generalize across individual participants, consistent with (Reiss & Stricker, 2012). In each iteration, one participant was held out for testing while the remaining 17 were split into training (13 participants) and validation (4 participants) sets. As a consequence, a different model was trained for each held-out participant, resulting in 18 independent models.

Several performance metrics, such as accuracy, Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), and both macro and class-specific F1-scores, were computed for each of the 18 folds and summarized as mean values with 95% confidence intervals (Altman, 1990). Additionally, to assess the overall performance across the whole dataset, a pooled confusion matrix was created by merging the predictions and ground truth labels from all 18 subject-specific test sets.

To improve temporal consistency of subsequent predictions and reduce the impact of instantaneous classification noise, a post-processing refinement classification strategy based on Time Majority Voting (TMV) was applied, consistent with the approach proposed in (Dou et al., 2022). Specifically, TMV exploits the time continuity of the IMU signals to ensemble the model predictions over the 67% overlapping 3-second windows. In fact, because each 1-second interval is observed by three different windows, the final TMV label is assigned based on the highest mean probability computed across all three. To evaluate the effectiveness of TMV classification, the resulting performance metrics and pooled confusion matrices were systematically compared against those obtained using a single-window classification strategy.

Finally, an exploratory dimensionality reduction analysis was run on the features extracted after applying the self-attention pooling layer in order to qualitatively assess the structure of the learned representations. The extracted features were reduced to 50 through a principal component analysis (Kurita, 2020) before being embedded into a two-dimensional space by means of the t-distributed Stochastic Neighbor Embedding (t-SNE) approach (van der Maaten & Hinton, 2008).

To justify the use of a DL approach, the proposed model was benchmarked against a standard Machine Learning baseline based on a Random Forest classifier trained on handcrafted features (see Appendix C).

3. Results

3.1. Ablation study

The results of the ablation study are reported in Table 1. The two top-performing configurations were *Acc+Gyro+Mag* and *Acc+Mag*, with no statistically significant difference in median macro F1-score ($p = 0.532$), while all other sensor combinations showed significantly lower performance ($p < 0.05$). Interestingly, among the single-sensor configurations, the *Acc*-only setup resulted as the most informative, significantly outperforming both *Gyro* ($p = 0.035$) and *Mag* ($p < 0.001$).

Although the two top-performing configurations showed comparable macro F1-scores, *Acc+Gyro+Mag* reached the highest median value. Furthermore, a qualitative analysis of the class-specific recall metrics highlighted that *Acc+Gyro+Mag* improved, still without statistical significance, the median recall value for both Rest (0.730 vs 0.664 for *Acc+Mag*) and Balance (0.801 vs 0.764 for *Acc+Mag*). As a consequence, the complete IMU sensors setup was selected as the optimal configuration for further studies.

Table 1. Ablation Study results for different sensor configurations, ranked in descending order by F1 Macro. Values are reported as median ± interquartile range (IQR). Configurations with different letters in the Compact Letter Display (CLD) are significantly different ($p < 0.05$).

SENSORS CONFIGURATION	F1 MACRO	CLD
Acc+Gyro+Mag	0.800 ± 0.066	A
Acc+Mag	0.795 ± 0.055	A
Gyro+Mag	0.762 ± 0.079	B
Acc	0.719 ± 0.047	B
Acc+Gyro	0.719 ± 0.069	B
Gyro	0.715 ± 0.050	C
Mag	0.654 ± 0.053	D

3.2. Model validation and deployment

Table 2 summarizes the mean and confidence intervals of the rehabilitation activity classification model performance metrics across the 18 participants, according to the LOSO

cross-validation approach (see also Appendix B). The system proved high generalization capability, with the performance metrics having relatively small IQR around the median value. The application of the TMV classification strategy led to a small, yet consistent improvement ranging between $+0.52\%$ and $+1.42\%$ across almost all performance metrics compared to the single-window strategy. Furthermore, the proposed ResNet model consistently outperformed the Random Forest baseline. As reported in Appendix C, the most significant improvements were observed within the pediatric patient group, particularly for Rest (+4.3% improvement, $p = 0.038$) and Balance (+5.2% improvement, $p = 0.004$).

However, classification performance varied across the rehabilitation activities. The system predicted Walk with near-perfect accuracy, reaching a median F1 score of 0.965, while it registered more difficulty in correctly predicting both Rest and Balance activities. Accordingly, the pooled confusion matrices reported in Figure 2 clearly reveal that the vast majority of misclassifications performed by the system occurred between Rest and Balance, while Walk was rarely confused with the other two activities, regardless of the classification strategy. In fact, Rest misclassifications were attributed to Balance in 93.5% of cases under the single-window classification strategy, and in 94% of cases using TMV, while Balance was misclassified as Rest in 90% and 91.7% of cases, respectively.

The results of the exploratory t-SNE analysis are displayed in Figure 3. The two-dimensional projection of the latent feature space revealed a unique topological organization of the samples. The Walk samples generated several distinct, well-separated clusters that stood apart from the other activities samples. Conversely, the Rest and Balance samples exhibited a less clear distribution, often appearing in close proximity or even overlapped in some clusters.

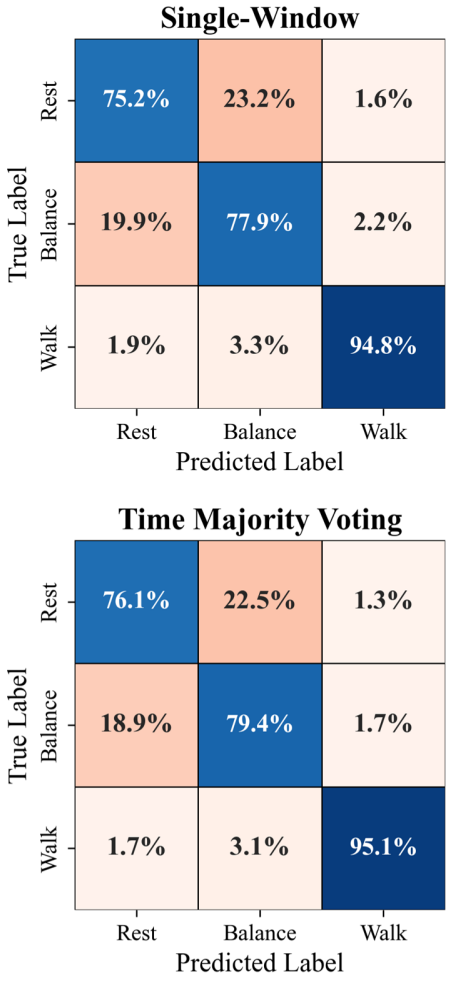

Figure 2. Row-wise normalized pooled confusion matrices. The top panel shows the results for the Single-Window classification strategy, while the bottom panel illustrates the performance after applying Time Majority Voting.

3.3. Runtime performance

To quantitatively assess the feasibility of a real-time deployment of the proposed ResNet model, a runtime performance analysis was conducted using the same PC hardware employed for model training. Specifically, 100,000 model inferences on 3-second windows were simulated and the execution time for each forward pass was collected. The resulting average inference time was 1.6 milliseconds, with a maximum recorded inference time of approximately 11 milliseconds.

4. Discussion

The current study presented a deep learning-based framework for the automated classification of rehabilitation activities, which was intended to work in a mixed pediatric cohort consisting of both neurotypical subjects and patients with neuromotor disorders. The main objective was to develop a lightweight solution that could generalize across multiple motor and cognitive functional domains without requiring any further subject-specific calibration or the use of transfer learning approaches.

The use of the LOSO validation approach was essential for ensuring a robust and realistic estimation of the model performance. In fact, according to (Bragança et al., 2022), standard k-fold cross-validation often leads to performance overestimation in HAR due to data leakage; on the other hand, the subject-independent approach used in this study stressed the model capability to generalize to unseen subjects, even though belonging to different motor functional domains. This robustness was particularly relevant given the mixed nature of the dataset, which introduced non-negligible vari-

Table 2. LOSO Cross-Validation ($N = 18$) performance metrics summary for both single-window and Time Majority Voting (TMV) classification strategies. Values are reported as mean [95% confidence interval]. The Δ Improvement represents the relative percentage change in performance moving from the single-window to the TMV classification strategies. Abbreviation: ROC AUC (Receiver Operating Characteristic Area Under the Curve).

METRIC	SINGLE-WINDOW	TIME MAJORITY VOTING	Δ IMPROVEMENT
ACCURACY	0.841 [0.810, 0.871]	0.848 [0.818, 0.879]	+0.83%
ROC AUC	0.943 [0.926, 0.961]	0.943 [0.926, 0.961]	-
F1 (MACRO)	0.807 [0.767, 0.846]	0.815 [0.775, 0.855]	+0.99%
F1 (REST)	0.706 [0.624, 0.787]	0.716 [0.634, 0.798]	+1.42%
F1 (BALANCE)	0.749 [0.698, 0.799]	0.759 [0.706, 0.812]	+1.34%
F1 (WALK)	0.965 [0.956, 0.974]	0.970 [0.961, 0.978]	+0.52%

ability in motor patterns across participants. This inter-subject variability was visually confirmed by the subject-wise t-SNE analysis provided in Appendix E, which highlighted the presence of distinct motor fingerprints, as suggested by (Manuello et al., 2025), characterized by subject-specific cluster distributions in the latent space. Despite this, the model successfully generalized, allowing to address a critical challenge identified in (Khaked et al., 2025), who observed that deep learning models trained on limited lab-controlled data tend to obtain lower performance when tested on data belonging to different domains compared to the training set. To further investigate this limitation and test the absolute boundaries of the proposed ResNet model's generalization capabilities beyond the pediatric domain, an exploratory zero-shot evaluation on an independent, open-access dataset of healthy adults (*RealWorld HAR* (Sztyler & Stuckenschmidt, 2016)) was carried out. As detailed in Appendix D, the ResNet model reported exceptional performance in classifying both Rest and Walk activities in a out-of-distribution scenario without requiring any further training or fine-tuning. Besides that, the overall achieved macro F1-score aligned with similar works, as in (Tørring et al., 2024), where the authors also reported robust subject-independent performance using a dataset containing data from both typically developing children and pediatric patients with cerebral palsy.

In terms of activity-specific performance, the model allowed for a nearly-perfect classification of the Walk activity, proving that the encoder and the self-attention pooling layer were able to extract very distinctive features that consistently represented the unique kinematic patterns associated with gait, both in neurotypical and neuromotor disorder participants. On the other hand, the ResNet-based model struggled more in providing to the classification head clearly distinct and representative latent features when dealing with Rest and Balance, thus leading to a non-negligible cross-confusion between the two low-dynamic activities. There were two major factors that contributed to this. From a methodological point of view, the Rest phase was designed to maximize ecological validity of the daily clinical practice that the study attempted to observe, and it worked as a silent subject profiling phase rather than a controlled task. In fact, all participants were asked to remain calm and relaxed, but they were not required to maintain a rigid posture, nor were they restricted from sitting, standing, or moving freely within the environment while the procedures prior to the beginning of the actual rehabilitation exercises were completed; as a result, the Rest activity included behaviors that introduced variability into the IMU data. From a biomechanical point of view, instead, the kinematic patterns of passive standing (common in Rest) and static, but active, balance exercises (common in Balance) are highly correlated and, as a consequence, the extracted latent features were partially overlapped. The difficulty in separating passive from active standing was consistent with the literature. Specifically, similar challenges were reported in (Kowalsky et al., 2024), where it was noted that machine learning-based prediction models refinement was essential to distinguish between passive and active standing.

The uncertainty between Rest and Balance activities was also evident when analyzing the latent feature space using the t-SNE approach in Figure 3. In fact, the two-dimensional t-SNE projection of the latent feature space extracted from all samples highlighted that while Walk samples formed well-separated clusters, the representations for Rest and Balance were much closer to each other and, sometimes, overlapped, thus suggesting that the encoder extracted latent features that were highly correlated for these two activities. In addition, results from the ablation study shed light on how different sensor configurations compensated for this uncertainty. Specifically, although the accelerometer was the most informative sensor when used alone, the integration of the magnetometer was essential to maximize performance for Rest and Balance activities. This finding was partially in contrast with some literature. Specifically, while in (Shang et al., 2022) the authors highlighted the added value of the gyroscopic signal in various HAR tasks, this study might suggest that, at least for low-dynamic rehabilitation tasks, the absolute orientation obtained through the magnetometer provided much more information than angular velocity.

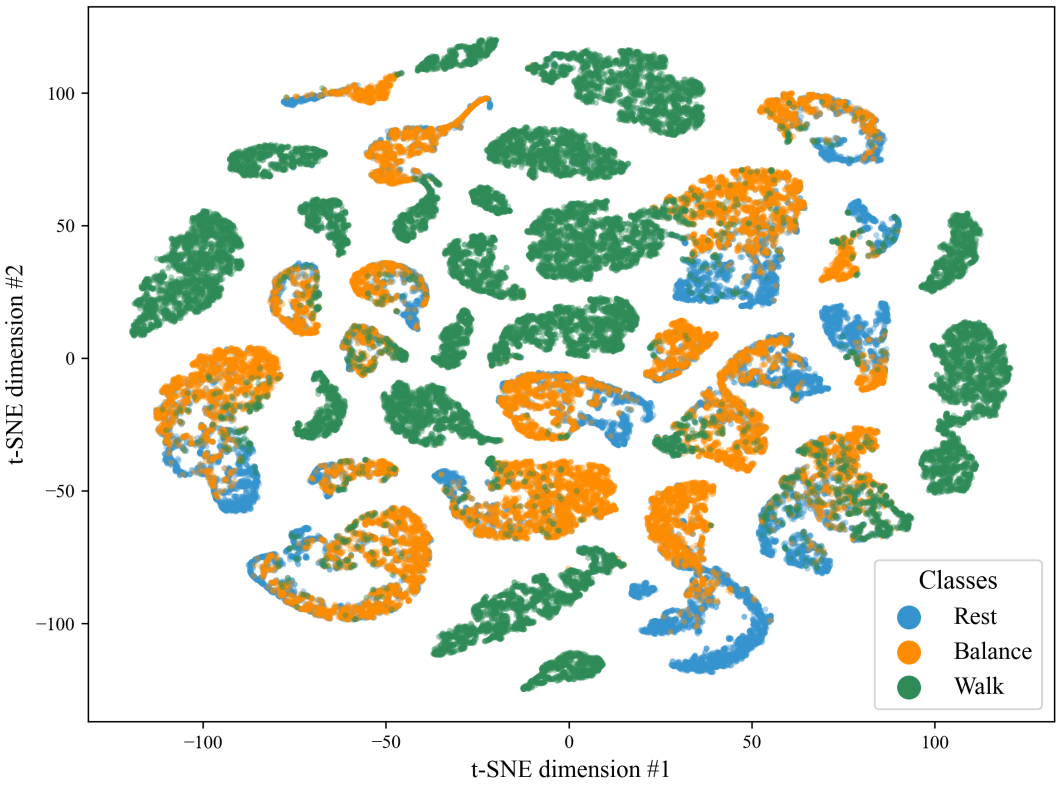

Figure 3. Two-dimensional t-SNE projection of the high-dimensional latent feature space. Each point represents a dataset sample and is color-coded by rehabilitation activity (Rest vs Balance vs Walk).

Finally, the proposed model architecture proved clear advantages for deployment in a clinical motor rehabilitation setting. The inclusion of a self-attention pooling layer allowed the model to dynamically weight different temporal segments, which made it easier to capture long-term dependencies in more complex activities. Furthermore, the model was intended to be computationally efficient, in line with the trend towards lightweight architectures, as in (Mekruksavanich et al., 2025), that facilitate the model implementation on edge devices. The combination of the lightweight design, which guarantees inference times in the order of milliseconds on standard hardware, the minimally intrusive L5 sensor placement, and the increased stability provided by the Time Majority Voting classification strategy, demonstrated the system's suitability for real-time rehabilitation activity monitoring.

5. Conclusions

The current study presented and validated a lightweight deep residual network for the automated classification of rehabilitation activities in a heterogeneous pediatric population. Using a single IMU sensor placed at the L5 vertebra, the system consistently generalized across varying motor functional domains, effectively tackling the domain shift challenge without subject-specific calibration. Specifically, while the gait activity was recognized with near-perfect sensitivity, the model also showed its potential ability to distinguish between rest and active standing. As a result, the proposed framework provides a significant step toward the use of ecological, unobtrusive tools for the objective recognition and quantification of rehabilitation activity in a clinical context.

Acknowledgments

This work was funded by the Italian Ministry of Health (Ricerca Corrente 2025-2026 awarded to Dr. E. Biffi), and by the Italian Ministry of University and Research (Doctoral Scholarship awarded to S. Costantini). The authors would like to deeply thank all the children and their families for their participation, patience, and commitment to this study. Furthermore, we extend our sincere gratitude to the therapists of the Scientific Institute IRCCS "E. Medea" who conducted the rehabilitation sessions and supported the data collection process.

Limitations

The study is certainly not exempt from limitations. First, the sample size was relatively small ($N = 18$), with the pathological cohort consisting of only eight subjects with heterogeneous neuromotor disorders. While the LOSO validation proved the model generalization capabilities across different motor functional domains, a larger and more stratified dataset would be required to definitively confirm its robustness across specific pathology subtypes and severity levels. Second, the ecological definition of the Rest class introduced some ambiguity. Since participants were not restricted to a rigid posture but allowed to move freely, the biomechanical boundary between Rest and Balance exercises was occasionally inconsistent, thus limiting the classification performance for the low-dynamic activities. To address these challenges, future extensions of this work will include larger cohorts and focus on a finer-grained classification within these rehabilitation macro-activities, such as distinguishing between dual-task exercises, maximal effort tasks, and specific motor control tasks. Finally, data collection was limited to a single or few sessions per subject. As a result, the current study addresses the system's instantaneous classification performance but does not test its longitudinal stability over time or its sensitivity to kinematic changes due to a full rehabilitation cycle with the GRAIL system.

Ethical Statement

The study was performed in accordance with the Declaration of Helsinki, and was approved by the Ethical Committee Lombardia 2 (protocol code: L2-246; date of approval: March 22nd, 2025). The study is registered at ClinicalTrials.gov (identifier: NCT06993389). Participants' guardians signed a written informed consent.

All experiments were conducted on a workstation equipped with an Intel Core i7 CPU and an NVIDIA GeForce GTX 1650 Ti GPU. As a consequence, the CO2 emissions and energy consumption associated with this study are negligible compared to large-scale deep learning models training.

The development of the current work relied on open-source software libraries (Python, PyTorch, SciPy, scikit-learn), which were used in full compliance with their respective licenses.

Finally, the authors acknowledge the use of Large Language Models (Gemini 1.5 Pro, Google) exclusively for the purpose of refining the linguistic quality of the manuscript. The scientific content, data analysis, and conclusions remain the original work of the authors, who reviewed and double-checked all AI-generated output.

References

Ahmadi, M. N., O'Neil, M. E., Baque, E., Boyd, R. N., and Trost, S. G. Machine learning to quantify physical activity in children with cerebral palsy: Comparison of group, group-personalized, and fully-personalized activity classification models. *Sensors*, 20(14):3976, July 2020. ISSN 1424-8220. doi: 10.3390/s20143976. URL http://dx.doi.org/10.3390/s20143976.

Altman, D. G. *Practical Statistics for Medical Research*. Chapman and Hall/CRC, November 1990. ISBN 9780429258589. doi: 10.1201/9780429258589. URL http://dx.doi.org/10.1201/9780429258589.

Bragança, H., Colonna, J. G., Oliveira, H. A. B. F., and Souto, E. How validation methodology influences human activity recognition mobile systems. *Sensors*, 22 (6):2360, March 2022. ISSN 1424-8220. doi: 10.3390/s22062360. URL http://dx.doi.org/10.3390/s22062360.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001. doi: 10.1023/A:1010933404324.

Cao, J., Xie, S. Q., Das, R., and Zhu, G. L. Control strategies for effective robot assisted gait rehabilitation: The state of art and future prospects. *Medical Engineering; Physics*, 36(12):1555–1566, December 2014. ISSN 1350-4533. doi: 10.1016/j.medengphy.2014.08.005. URL http://dx.doi.org/10.1016/j.medengphy.2014.08.005.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(4):1–40, May 2021. ISSN 1557-7341. doi: 10.1145/3447744. URL http://dx.doi.org/10.1145/3447744.

Cope, S. and Mohn-Johnsen, S. The effects of dosage time and frequency on motor outcomes in children with cerebral palsy: A systematic review. *Developmental Neurorehabilitation*, 20(6):376–387, February 2017. ISSN 1751-8431. doi: 10.1080/17518423.2017.1282053. URL http://dx.doi.org/10.1080/17518423.2017.1282053.

Dou, G., Zhou, Z., and Qu, X. *Time Majority Voting, a PC-Based EEG Classifier for Non-expert Users*, pp. 415–428. Springer Nature Switzerland, 2022. ISBN 9783031176180. doi: 10.1007/978-3-031-17618-0_29. URL http://dx.doi.org/10.1007/978-3-031-17618-0_29.

Fundarò, C., Giardini, A., Maestri, R., Traversoni, S., Bartolo, M., and Casale, R. Motor and psychosocial impact of robot-assisted gait training in a real-world rehabilitation setting: A pilot study. *PLOS ONE*, 13(2): e0191894, February 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0191894. URL http://dx.doi.org/10.1371/journal.pone.0191894.

González Barral, C. and Servais, L. Wearable sensors in paediatric neurology. *Developmental Medicine; Child Neurology*, 67(7):834–853, January 2025. ISSN 1469-8749. doi: 10.1111/dmcn.16239. URL http://dx.doi.org/10.1111/dmcn.16239.

Guadagnoli, M. A. and Lee, T. D. Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior*, 36(2):212–224, July 2004. ISSN 1940-1027. doi: 10.3200/jmbr.36.2.212-224. URL http://dx.doi.org/10.3200/JMBR.36.2.212-224.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, June 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

Kańtoch, E. Human activity recognition for physical rehabilitation using wearable sensors fusion and artificial neural networks. In *2017 Computing in Cardiology (CinC)*, pp. 1–4, 2017. doi: 10.22489/CinC.2017.296-332.

Khaked, A. A., Oishi, N., Roggen, D., and Lago, P. In shift and in variance: Assessing the robustness of har deep learning models against variability. *Sensors*, 25 (2):430, January 2025. ISSN 1424-8220. doi: 10.3390/s25020430. URL http://dx.doi.org/10.3390/s25020430.

Kim, Y.-W., Joa, K.-L., Jeong, H.-Y., and Lee, S. Wearable imu-based human activity recognition algorithm for clinical balance assessment using 1d-cnn and gru ensemble model. *Sensors*, 21(22):7628, November 2021. ISSN 1424-8220. doi: 10.3390/s21227628. URL http://dx.doi.org/10.3390/s21227628.

Kowalsky, R., van Werkhoven, H., Meucci, M., Quinn, T., Stoner, L., Hearon, C., and Gibbs, B. B. Distinguishing passive and active standing behaviors from accelerometry. *Journal for the Measurement of Physical Behaviour,*, 7 (1), 2024.

Kurita, T. *Principal Component Analysis (PCA)*, pp. 1–4. Springer International Publishing, 2020. ISBN 9783030032432. doi: 10.1007/978-3-030-03243-2_649-1. URL http://dx.doi.org/10.1007/978-3-030-03243-2_649-1.

Manuello, J., Ciceri, T., Longatelli, V., Maronati, C., Biffi, E., Cavallo, A., and Casartelli, L. Mapping the complexity of motor variability: From individual space of variability to motor fingerprints. *Behavior Research Methods*, 57(5), April 2025. ISSN 1554-3528. doi: 10.3758/s13428-025-02635-0. URL http://dx.doi.org/10.3758/s13428-025-02635-0.

Mekruksavanich, S., Jitpattanakul, A., Sitthithakerngkiet, K., Youplao, P., and Yupapin, P. Resnet-se: Channel attention-based deep residual network for complex activity recognition using wrist-worn wearable sensors. *IEEE Access*, 10:51142–51154, 2022. ISSN 2169-3536. doi: 10.1109/access.2022.3174124. URL http://dx.doi.org/10.1109/ACCESS.2022.3174124.

Mekruksavanich, S., Tancharoen, D., and Jitpattanakul, A. Efficient and lightweight human activity recognition based on wearable sensors using light residual neural network. In *2025 24th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 316–321. IEEE, October 2025. doi: 10.1109/iscit67082.2025.11231512. URL http://dx.doi.org/10.1109/ISCIT67082.2025.11231512.

Novak, I., Morgan, C., Fahey, M., Finch-Edmondson, M., Galea, C., Hines, A., Langdon, K., Namara, M. M., Paton, M. C., Popat, H., Shore, B., Khamis, A., Stanton, E., Finemore, O. P., Tricks, A., te Velde, A., Dark, L., Morton, N., and Badawi, N. State of the evidence traffic lights 2019: Systematic review of interventions for preventing and treating children with cerebral palsy. *Current Neurology and Neuroscience Reports*, 20(2), February 2020. ISSN 1534-6293. doi: 10.1007/s11910-020-1022-z. URL http://dx.doi.org/10.1007/s11910-020-1022-z.

Oleh, U., Obermaisser, R., and Ahammed, A. S. A review of recent techniques for human activity recognition: Multimodality, reinforcement learning, and language models. *Algorithms*, 17(10):434, September 2024. ISSN 1999-4893. doi: 10.3390/a17100434. URL http://dx.doi.org/10.3390/a17100434.

Reiss, A. and Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*. IEEE, June 2012. doi: 10.1109/iswc.2012.13. URL http://dx.doi.org/10.1109/ISWC.2012.13.

Shang, M., De Raedt, W., Varon, C., and Vanrumste, B. Are gyroscopes an added value in leave-one-subject-out activity recognition with imus? In *2022 44th Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)*, pp. 2399–2402. IEEE, July 2022. doi: 10.1109/embc48229.2022.

9871845. URL http://dx.doi.org/10.1109/EMBC48229.2022.9871845.

Steinmetz, J. D., Vos, T., and Dua, T. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Neurology*, 23(4):344–381, April 2024. ISSN 1474-4422. doi: 10.1016/s1474-4422(24)00038-3. URL http://dx.doi.org/10.1016/S1474-4422(24)00038-3.

Sztyler, T. and Stuckenschmidt, H. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–9. IEEE, March 2016. doi: 10.1109/percom.2016.7456521. URL http://dx.doi.org/10.1109/PERCOM.2016.7456521.

Tørring, M. F., Logacjov, A., Brændvik, S. M., Ustad, A., Roeleveld, K., and Bardal, E. M. Validation of two novel human activity recognition models for typically developing children and children with cerebral palsy. *PLOS ONE*, 19(9):e0308853, September 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0308853. URL http://dx.doi.org/10.1371/journal.pone.0308853.

van der Maaten, L. and Hinton, G. E. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://api.semanticscholar.org/CorpusID:5855042.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Welch, P. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, June 1967. ISSN 0018-9278. doi: 10.1109/tau.1967.1161901. URL http://dx.doi.org/10.1109/TAU.1967.1161901.

A. Demographics and clinical characteristics of the participants

Table 3. Demographic and clinical characteristics of each subject. Abbreviations: INT (Intervention Group), CTX (Control Group), ABI (Acquired Brain Injury), CP (Cerebral Palsy), GMFCS (Gross Motor Function Classification System).

SUBJECT	AGE	GENDER	HEIGHT (CM)	WEIGHT (KG)	GROUP	DIAGNOSIS	GMFCS
1	9	F	125	27.5	INT	ABI	I
2	11	F	145	49.0	INT	CP	II
3	7	F	127	27.5	INT	ABI	I
4	17	M	155	47.0	INT	CP	II
5	6	F	128	28.0	CTX	-	-
6	8	M	150	40.0	CTX	-	-
7	12	M	154	45.0	CTX	-	-
8	13	M	147	39.5	INT	CP	II
9	11	F	167	48.0	CTX	-	-
10	8	M	122	21.0	CTX	-	-
11	6	M	118	18.0	CTX	-	-
12	13	M	150	40.0	CTX	-	-
13	9	F	120	29.0	INT	ABI	I
14	9	F	141	28.0	CTX	-	-
15	6	F	115	19.0	CTX	-	-
16	9	M	125	22.0	INT	CP	I
17	8	F	139	31.0	CTX	-	-
18	8	M	138	31.0	INT	ABI	I

B. Detailed Leave-One-Subject-Out results

Table 4. Subject-wise performance metrics from the Leave-One-Subject-Out Cross-Validation ($N = 18$). The table reports Accuracy, Receiver Operating Characteristic Area Under the Curve (ROC AUC), Macro F1, and class-specific F1 scores for Rest, Balance, and Walk activities for each subject.

SUBJECT	ACCURACY	ROC AUC	MACRO F1	F1 (REST)	F1 (BALANCE)	F1 (WALK)
1	0.853	0.954	0.836	0.772	0.787	0.948
2	0.767	0.926	0.746	0.646	0.629	0.963
3	0.871	0.961	0.855	0.788	0.806	0.970
4	0.826	0.943	0.793	0.720	0.695	0.965
5	0.785	0.899	0.737	0.434	0.802	0.976
6	0.798	0.923	0.768	0.647	0.695	0.962
7	0.935	0.988	0.923	0.913	0.892	0.963
8	0.865	0.962	0.850	0.791	0.790	0.970
9	0.820	0.938	0.823	0.792	0.719	0.957
10	0.729	0.855	0.690	0.383	0.715	0.974
11	0.847	0.957	0.830	0.757	0.768	0.963
12	0.757	0.892	0.617	0.400	0.512	0.938
13	0.934	0.987	0.920	0.912	0.849	0.999
14	0.829	0.941	0.775	0.714	0.633	0.978
15	0.915	0.976	0.889	0.802	0.870	0.995
16	0.874	0.938	0.817	0.606	0.891	0.953
17	0.917	0.989	0.887	0.913	0.770	0.977
18	0.809	0.945	0.765	0.715	0.657	0.924

C. Machine Learning Baseline Benchmark

To justify the implementation of a deep residual network, its performance was benchmarked against a standard ML approach. Specifically, a Random Forest (RF) classifier (Breiman, 2001) using the same LOSO cross-validation framework described in Section 2.6 was exploited.

For the RF classifier, eight handcrafted features were extracted from each of the nine IMU channels for every 3-second window, resulting in a feature vector of 72 entries per sample. The extracted features consisted of time-domain statistical metrics (mean, median, standard deviation, inter-quartile range, minimum, and maximum) and frequency-domain metrics (peak frequency and total spectral energy) computed via Welch's power spectral density approach (Welch, 1967). The RF model was initialized with 50 estimators and balanced class weights to handle dataset unbalance.

The proposed ResNet model's performance was compared against the RF baseline across all subjects ($N = 18$), the neurotypical control group ($N = 10$), and the intervention group with neuromotor disorders ($N = 8$). Statistical significance was tested using a one-tailed Wilcoxon signed-rank test.

As reported in Table 5, the ResNet architecture consistently outperformed the RF baseline. Notably, statistically significant improvements were observed in the intervention group, demonstrating that the Deep Learning model is significantly more robust in handling the complex kinematic variability introduced by neuromotor disorders, particularly for critical activities such as Rest (+4.3%, $p = 0.038$) and Balance (+5.2%, $p = 0.004$).

Table 5. Performance comparison between the Random Forest (RF) baseline and the proposed ResNet model. Values are reported for all subjects (All), the neurotypical control group (CTX), and the intervention group (INT). Significant differences are highlighted in bold.

METRICS	GROUP	RF	RESNET	Δ IMPROVEMENT	P-VALUE
F1 MACRO	ALL	0.794	0.807	1.6%	0.120
	CTX	0.792	0.794	0.3%	0.577
	INT	**0.796**	**0.823**	**3.4%**	**0.007**
F1 (REST)	ALL	0.692	0.706	2.0%	0.210
	CTX	0.675	0.676	0.1%	0.577
	INT	**0.713**	**0.744**	**4.3%**	**0.038**
F1 (BALANCE)	ALL	0.727	0.749	3.0%	0.059
	CTX	0.729	0.737	1.1%	0.348
	INT	**0.725**	**0.763**	**5.2%**	**0.004**
F1 (WALK)	ALL	0.956	0.965	0.9%	0.180
	CTX	0.965	0.968	0.3%	0.577
	INT	0.953	0.968	1.5%	0.138

D. Zero-shot generalization on out-of-distribution data

To extensively evaluate the generalization capabilities of the proposed model beyond the population properties observed in the study pediatric cohort, a zero-shot evaluation (i.e., inference without any fine-tuning or retraining) on the independent, open-access *RealWorld HAR* dataset (Sztyler & Stuckenschmidt, 2016) was conducted.

This specific dataset was chosen to test the model against drastic out-of-distribution data. The *RealWorld HAR* dataset contains data collected from 15 healthy adults (mean age 31.9 ± 12.4 years, mean height 173.1 ± 6.9 cm, mean weight 74.1 ± 13.8 kg, eight males and seven females), a population that is biomechanically and demographically different from the current study cohort. Furthermore, IMU data was acquired using the built-in sensors of smartphones sampled at 50 Hz, rather than professional IMU sensors sampled at 128 Hz.

Despite these differences, the *RealWorld HAR* dataset contains 9-axis IMU data collected from the subjects' waist, a placement biomechanically equivalent to the L5-vertebra location used throughout this study experimental campaign. Additionally, the dataset records daily-life activities that can reasonably be associated with the study target activities. Specifically, the *Sitting* and *Standing* activities of the *RealWorld HAR* dataset were mapped to Rest, while the *Walking* activity was mapped to Walk. Unfortunately, a suitable match for the Balance activity could not be identified. In fact, within the current protocol, this activity represents an active, dynamic task characterized by continuous postural micro-adjustments, rendering it incompatible with the passive *Standing* activity of the *RealWorld HAR* dataset.

To perform the zero-shot evaluation, the *RealWorld HAR* data was pre-processed to match the input requirements of the proposed ResNet model. The raw accelerometer, gyroscope, and magnetometer signals were spatially aligned to the L5 sensor orientation, temporally synchronized, and upsampled to 128 Hz using linear interpolation. Subsequently, the data underwent the same conditioning pipeline applied in this study (see Section 2.3) before being segmented into 3-second

sliding windows.

The pre-trained ResNet model was fed with this unseen, out-of-distribution data, achieving outstanding zero-shot F1-scores of 0.963 on Rest and 0.973 on Walk, respectively, on average across the 15 unseen subjects. These results proved that the proposed model successfully learned generalized, universal kinematic features that are not strictly confined to the pediatric domain, allowing for robust deployment in compatible sensor setups without the need for retraining.

E. Subject-wise latent feature distribution according to the t-SNE analysis

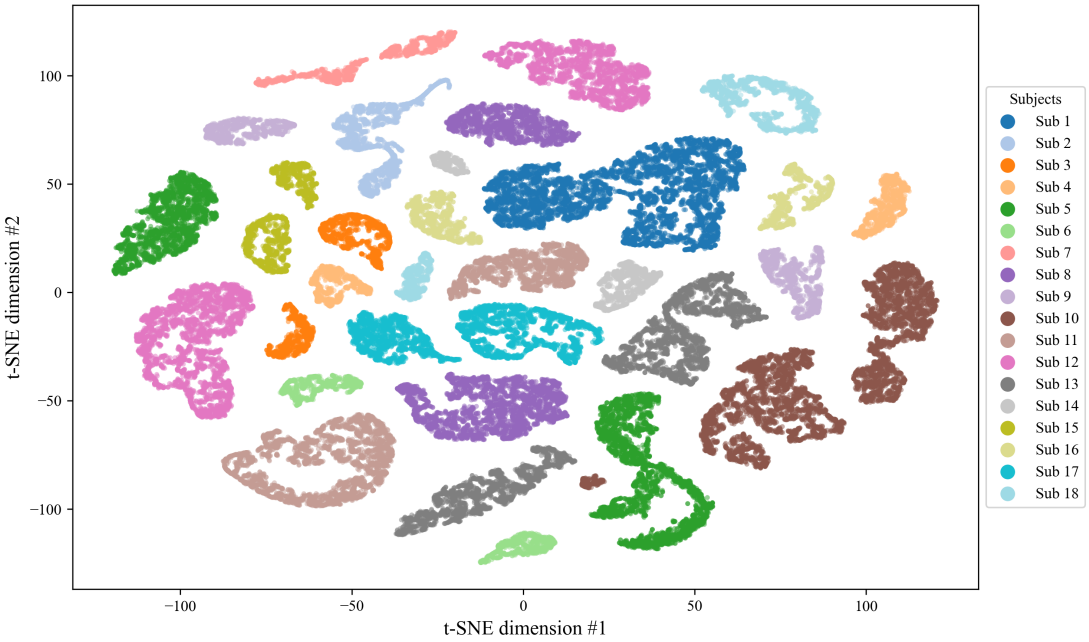

Figure 4. Two-dimensional t-SNE projection of the high-dimensional latent feature space, colored by subject ID.