Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

A Study on Predicting Adolescent Cross-Sport Performance Using Shared Feature Space and Multi-Task Regression

Yue Hou [1,*]

[1] College of Engineering, Northeastern University, Boston, MA, 02115, United States
* Correspondence: Yue Hou, College of Engineering, Northeastern University, Boston, MA, 02115, United States

Abstract: Common transfer patterns exist across different sports disciplines in adolescents regarding abilities such as speed, explosiveness, and rhythm control. This study constructs a cross-discipline prediction framework based on shared feature space, targeting throwing velocity, sprint performance, and jumping ability. The framework comprises a general ability feature layer, discipline-specific feature layer, and multi-task regression head. General features include motion rhythm statistics, velocity distribution, acceleration trends, and training load indicators, while the discipline layer incorporates discipline-specific motion information. Predictions are performed using multi-head LightGBM. Experiments on 2.1 million sequence data points from 619 athletes demonstrate an average MAPE of 6.5%, representing a 22.7% improvement over single-task models. The framework maintains stable advantages even under sparse data conditions, indicating that the shared feature structure effectively enhances cross-sport prediction capabilities for adolescent athletic performance.

Keywords: cross-sport prediction; multi-task learning; shared feature space; athletic performance; LightGBM

1. Introduction

Adolescence represents a critical period where athletic ability structures rapidly form alongside multidimensional development. The latent commonalities in physical characteristics and rhythm control patterns across different sports provide a theoretical foundation for constructing a unified performance prediction system. Practical training and talent selection processes often encounter challenges such as difficulty in cross-sport ability assessment and low validity of metric conversion, necessitating data-driven methods for ability transfer modeling. To address this, this paper proposes a predictive model integrating shared feature spaces with multi-task regression. It systematically combines multi-source high-frequency time-series data (e.g., IMU and GPS) with a LightGBM multi-head regression architecture to achieve joint modeling across multiple tasks such as throwing, sprinting, and jumping. The model design focuses on extracting universal abilities while compensating for task-specific variations, with the core objectives of enhancing prediction accuracy and cross-sport generalization capabilities. This research aims to provide an extensible data modeling paradigm for evaluating multi-event training in youth athletes, advancing athletic performance prediction systems toward multi-task collaboration and structured representation.

Received: 11 December 2025
Revised: 23 January 2026
Accepted: 06 February 2026
Published: 13 February 2026

2. Theoretical Foundations for Predicting Adolescent Multi-Event Performance

Youth athletic performance exhibits significant cross-event transferability, rooted in the common distribution of foundational motor abilities such as neuromuscular control, dynamic response, and rhythmic regulation. Within biokinesiology and multi-event training theory, metrics like speed, acceleration, and explosive power form transferable performance mappings across different motor skills [1]. The Multi-Task Learning (MTL) framework effectively captures these cross-sport capability correlations by simultaneously learning multiple prediction targets within a unified model. The Shared Feature Space theory emphasizes extracting stable, low-level physical movement patterns across multiple tasks to construct a unified capability representation layer, thereby enhancing the model's ability to model latent correlations between sports.

3. Design of a Cross-Sport Performance Prediction Model for Adolescents

3.1. Overall Model Architecture Design

The model adopts a shared-specific dual-path structure. The general ability feature layer extracts action rhythm statistics, velocity change rates, triaxial acceleration trends, and cumulative training load sampled at 0.1-second intervals. A sliding window reconstruction of 2.1 million sequences from 619 athletes forms a stable low-dimensional feature space [2]. The sport-specific feature layer integrates 45 dimensions of sport-unique pose angles, periodic rhythm factors, and local extrema points. The multi-task regression head constructs three LightGBM submodules to predict throwing velocity, 30m sprint time, and vertical jump height respectively. These submodules share upstream inputs with parameter settings of tree depth 7, learning rate 0.05, and subsampling rate 0.9. The overall model architecture is shown in Figure 1.

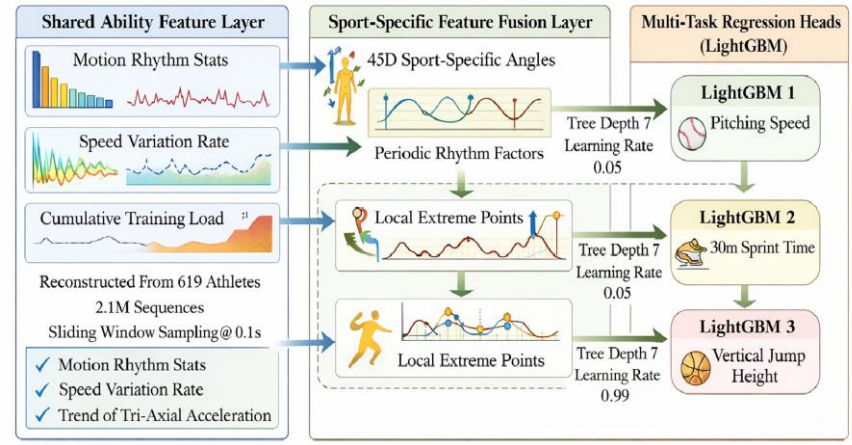

Figure 1. Architectural Diagram of the Youth Multi-Sport Performance Prediction Model.

3.2. Construction of the General Ability Feature Layer

The general ability feature layer was designed based on a multimodal sequence fusion strategy, focusing on low-order physical performance variables with stable transferability across sports disciplines. Data sources comprised 2.1 million temporal samples collected from 619 adolescent athletes using 3-axis IMUs and high-frequency GPS devices. Each sample has a duration of T = 4s and a sampling frequency of 50Hz. These samples are reorganized using a standard sliding window to form a fixed-dimensional input matrix $X \in \mathrm{IR}^{200\times d}$, where d = 18 represents the initial number of channels. These channels encompass four feature categories: velocity vector magnitude, rhythm period distribution, normalized acceleration rate of change, and unit-time training load [3]. The process of constructing the general ability expression tensor F_{shared} is defined by the following formula:

$$F_{shared} = \varphi\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha \cdot v_t + \beta \cdot \frac{da_t}{dt} + \gamma \cdot r_t + \delta \cdot l_t\right)\right) \quad (1)$$

Where v_t denotes the velocity magnitude vector at time t, $\frac{da_t}{dt}$ represents the acceleration change rate, r_t indicates the action rhythm statistical factor, l_t signifies the training load metric, $\alpha, \beta, \gamma, \delta \in$ IR is the feature weighting coefficient, and $\varphi(\cdot)$is the regularized activation mapping. This structure achieves high-dimensional redundant signal compression and cross-project representation extraction through temporal averaging and weighted filtering mechanisms, providing a structurally stable and semantically clear feature foundation for specificity compensation and multi-task modeling [4].

3.3. Project-Specific Feature Layer Design

The project-specific feature layer constructs a differentiated action representation matrix $F_{sport} \in \mathrm{IR}^{N \times d_s}$ after processing the shared feature base F_{shared}, where N = 2100000 represents the total number of sliding window samples and ds = 45 denotes the dedicated channel dimension for project actions [5]. The design objective is to provide fine-grained dynamic augmentation components for throwing, sprinting, and jumping tasks while maintaining input compatibility with the multi-task regressor. This layer employs a unified temporal alignment operator $A(\cdot)$and a local action encoder $\psi(\cdot)$to generate task inputsZ_i ($i \in \{1,2,3\}$ corresponds to the three regression tasks):

$$Z_i = \psi(A(F_{shared}, \mathrm{S}_{sport})), Z_i \in \mathrm{IR}^{200 \times d_s} \quad (2)$$

Where $S_{sport}\mathrm{IR}^{200 \times 3}$ represents the IMU three-axis angular velocity and attitude differential sequences aligned to a shared window via Dynamic Time Warping (DTW); $A(\cdot)$ denotes the temporal alignment fusion operator; and $\psi(\cdot)$ signifies the action encoding mapping adapted for LightGBM task inputs. To model the periodicity and velocity pattern variations of project actions, the action rhythm distribution R and velocity interval density function D are designed:

$$R = \frac{1}{K}\sum_{k=1}^{K} \frac{f_{cycle,k}}{200}, K = 619, f_{cycle,k} \in N \quad (3)$$

Where: K represents the number of athletes; $f_{cycle,k}$ denotes the number of detected peak main cycles within the 4-second window for the k th athlete; 200 corresponds to the number of sampling points in the 50Hz×4s window.

$$D(u) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{II}(v_n \in [u, u + \Delta u]), v_n \in \mathrm{IR}, \Delta u\text{=}0.5\mathrm{m/s} \quad (4)$$

Where v_n represents the velocity modulus within each window; $\mathrm{II}(\cdot)$denotes the indicator function; $[u, u + \Delta u]$ indicates the velocity binning at 0.5 m/s intervals; and $D(u)$ signifies the sample density within this velocity interval. The interface between the feature layer and the regressor is designed as a three-headed input projection $Z_i \rightarrow h_i(\cdot)$. The regressor is constructed using shared inputs and task-independent tree sets based on multi-head LightGBM, ensuring compatibility between the specific layer's output dimension and the regressor head parameters $depth$ = 7,lr=0.05,$subsample$=0.9.

3.4. Construction of Multi-Task Regression Head and Multi-Headed LightGBM Predictor

The multi-task regression head adopts a structurally parallel three-channel submodule architecture, constructing independent regression functions $f_i(\cdot)$ for each target task: pitching velocity, 30m sprint time, and vertical jump height. Inputs are derived from the fused feature tensor $Z_i \in \mathrm{IR}^{200 \times d_s}$, where ds=45 denotes the task-specific feature dimension, and $i \in \{1,2,3\}$ corresponds to the three tasks. Each submodule employs LightGBM to train a gradient-boosted tree ensemble comprising 200 base learners. A multi-task joint objective function based on mean squared error loss is designed as follows:

$$L_{MTL} = \sum_{i=1}^{3} w_i \cdot \frac{1}{n_i}\sum_{j=1}^{n_i} (\hat{y}_{ij} - y_{ij})^2 \quad (5)$$

Among these, y_{ij} denotes the true performance value of the j th sample in the i th task, $\hat{y}_{ij} = f_i(Z_{ij})$ represents the predicted value, w_i is the task weighting factor, and n_i indicates the number of samples for the corresponding task. Internally, the LightGBM model optimizes splitting complexity and residual fitting capability through structural regularization. Its weak learner construction process satisfies the following gain maximization criterion:

$$G(S) = \frac{1}{2}\left[\frac{\left(\sum_{j\in S_L} g_i\right)^2}{\sum_{j\in S_L} h_j+\lambda} + \frac{\left(\sum_{j\in S_R} g_i\right)^2}{\sum_{j\in S_R} h_j+\lambda} - \frac{\left(\sum_{j\in S} g_i\right)^2}{\sum_{j\in S} h_j+\lambda}\right] \quad (6)$$

Where S_L、S_R、S denote splitting into the left subset, right subset, and entire dataset, respectively; g_j and h_j represent the first-order and second-order derivatives of the sample, respectively; and λ is the leaf node regularization factor, set to 1.0. Each LightGBM regression head is configured with a depth of 7, learning rate of 0.05, subsampling rate of 0.9, and maximum leaf count capped at 32. This ensures the model maintains generalization capability while avoiding overfitting under high-dimensional, heterogeneous features [6]. To clearly illustrate the three adolescent athletic performance modeling objectives and their structural differentiation logic within the multi-task regression model, we visualize the modeling expressions for the three tasks-pitching velocity, 30m sprint time, and vertical jump height-by integrating actual data collection paths with simulation workflows (as shown in Figure 2).

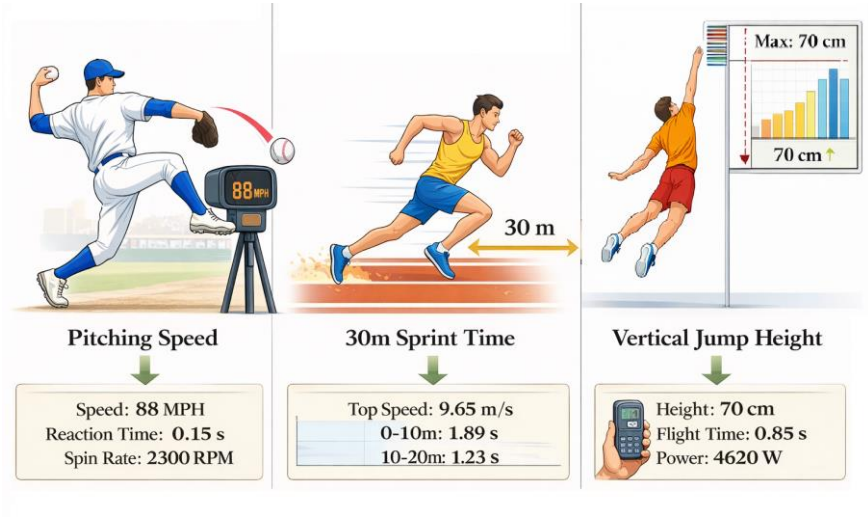

Figure 2. Simulation diagram of three adolescent athletic performance metrics.

3.5. Model Training and Parameter Settings

Model training employs a stratified allocation strategy, dividing athletes into training (70%), validation (15%), and test (15%) sets based on sport and gender proportions. All input features are standardized and concatenated using sliding windows to form tensor inputs $Z_i \in \mathrm{IR}^{200\times 45}$. The training process constructs an optimization objective based on L2 loss, incorporating regularization to control regressor complexity [7]. The loss function is defined as:

$$J(\theta) = \left\|Y - \hat{Y}\right\|_2^2 + \lambda\|\theta\|_2^2 \quad (7)$$

Where Y is the true label matrix, $\hat{Y}$ is the model prediction output, θ is the set of all regressor parameters, and λ is the L2 regularization coefficient (set to 0.01). LightGBM trained for 200 epochs per task with an early-stopping window of 30 epochs. The boosting strategy employed gradient boosting per tree, and training occurred on a single GPU with 32GB memory.

4. Experimental Results and Analysis

4.1. Experimental Setup and Evaluation Metrics

To validate the adaptability and stability of multi-task regression models in predicting cross-event athletic performance among adolescents, this experiment design fully integrates raw data with multi-dimensional metric requirements, with specific settings as follows: Data was selected from 619 adolescent athletes across throwing, sprinting, and jumping disciplines. The original dataset comprised 2.1 million sequences, each 4 seconds long at 50Hz sampling rate, containing IMU triaxial data, GPS velocity, and load information. After sliding window processing and standard normalization, the

data were proportionally divided into training (70%), validation (15%), and test (15%) sets, maintaining consistent gender and event type distributions; Evaluation metrics included Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Inter-Task Performance Variability. MAPE served as the primary evaluation criterion, measuring relative regression errors across disciplines for cross-target comparisons [8].

4.2. Baseline Model Comparison Analysis

For each of the three target tasks, separate LightGBM models were constructed as control groups, maintaining consistent input dimensions and dataset splits. Experimental results are shown in Table 1.

Table 1. Performance Comparison Between Single-Task and Multi-Task Models Across Different Tasks.

Model Structure	Prediction Task	MAPE (%)	MSE (units²)	Task-to-Task Volatility (%)
Single-Task Model	Pitching Speed	8.25	1.82	2.71
	30m Sprint Time	7.9	0.029	
	Vertical jump height	9.34	3.96	
Multi-task model	Pitching Speed	6.42	1.55	1.12
	30m Sprint Time	6.28	0.025	
	Jump height	6.8	3.31	

Table 1 shows that the multi-task model achieved an average MAPE of 6.50% across the three athletic performance predictions, representing a reduction of approximately 23.5% compared to the single-task model's 8.50%. This corresponds to an overall relative error reduction of 22.7%. Regarding the MSE metric, the values for throwing speed and jump height tasks decreased to 1.55 and 3.31, respectively, indicating significant advantages in both accuracy and stability. Concurrently, the inter-task performance variability decreased from 2.71% to 1.12%, further validating the coordinated generalization capability of the shared feature space structure in multi-objective modeling and providing a solid foundation for robustness testing under sparse sample conditions [9,10].

4.3. Performance Validation of Shared Feature Space

The effectiveness of the shared feature space was further validated by constructing a "de-shared" control group. While maintaining identical task structures and samples, the general capability layer was removed, and independent regression heads were trained solely using project-specific inputs. Results are shown in Figure 3.

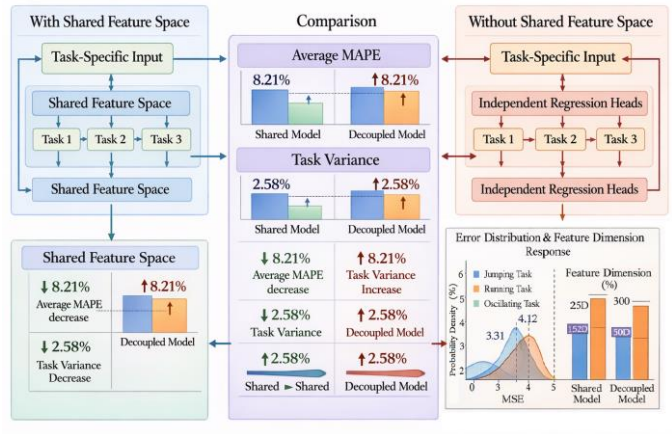

Figure 3. Structural diagram of model performance differences before and after introducing the shared feature space.

Experimental results show that the Mean Absolute Percentage Error (MAPE) increased to 8.21%, and the inter-task prediction volatility rose to 2.58%. This indicates that the absence of low-order general physical variables in the input features significantly weakens the model's stability and generalization ability for sequence prediction. Particularly in jump-type tasks, the Mean Squared Error (MSE) increased from 3.31 to 4.12, with a noticeable decline in model output accuracy, revealing insufficient adaptability to dynamically abrupt action patterns. Further analysis indicates that the absence of unified physical priors hinders the model's ability to form consistent cross-task abstractions, impeding the capture of generalizable patterns such as rhythm control and velocity changes within time series. Introducing the shared spatial structure effectively compressed frequently occurring noise redundancy in high-dimensional inputs, enhancing the compactness and distinctiveness of feature representations. This enabled the model to more stably extract common features across various tasks, strengthening its understanding of intrinsic patterns between different movement modes and its transfer capabilities. Overall, incorporating the sharing mechanism positively improved multi-task modeling performance, validating the critical value of universal physical information in sequence modeling.

4.4. Analysis of Transfer Effects Across Different Tasks

Under the shared feature space, asymmetric transfer pathways emerge between different motion tasks, revealing pronounced structural feature sharing. Results are shown in Figure 4.

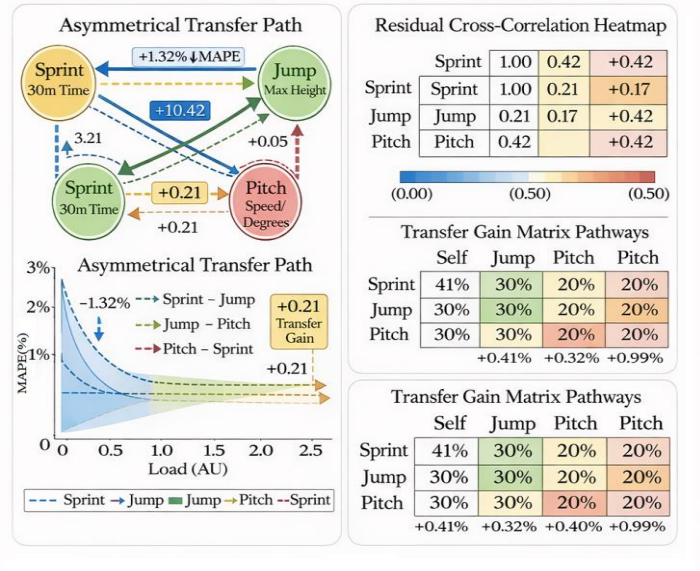

Figure 4. Task Transfer Results.

Results indicate that the sprint task exhibits positive transfer enhancement for the jump task. The MAPE of the 30m sprint time prediction model decreased by 1.32% after incorporating jump ability variables.The jump task had relatively limited influence on the throwing task, showing only minor error convergence during high-load training segments. Transfer residual analysis indicated the most significant rhythm-frequency coupling between throwing and sprinting, with residual cross-correlation rising to 0.42-significantly higher than throwing-jumping (0.17) and jumping-sprinting (0.21).Further construction of a three-dimensional transfer gain matrix revealed that the sprint task exhibited the strongest output contribution in feature sharing, while the jump task was most dependent on shared input representations, highlighting the dominant transfer role of speed and rhythm-related capability variables.

4.5. Robustness Testing in Sparse Sample Scenarios

The prediction stability of the multi-task model under sparse data conditions was evaluated by progressively reducing the sample size. When the training sample proportion decreased from 70% to 50%, the average MAPE across all three tasks increased from 6.50% to 7.12%, with the increase controlled within 0.62 percentage points. Further reduction to 30% samples increased the average MAPE to 7.84%, representing a 20.6% error increase relative to the full-sample condition. Corresponding MSE metrics showed: throwing speed task increased from 1.55 to 1.93, 30m sprint time from 0.025 to 0.031, and jump height from 3.31 to 4.02. Meanwhile, the inter-task performance variability remained stable at 1.86% under the 30% sample condition, showing no significant dispersion. Error growth rates varied across tasks: the jump task exhibited greater sensitivity to sample reduction, while the sprint task demonstrated relatively stable error variation. This reflects the shared feature space's robust support for fundamental motor ability representation under low sample density.

5. Conclusions

This research aims to provide an extensible data modeling paradigm for evaluating multi-event training in youth athletes, thereby advancing athletic performance prediction systems toward multi-task collaboration and structured representation. The proposed model is designed to extract universal athletic abilities while compensating for task-specific variations, with the core objectives of improving prediction accuracy and cross-sport generalization. The feature fusion mechanism effectively enhances model stability and generalization during multi-task modeling. General physical variables-represented by rhythm, speed, and acceleration-demonstrate critical value in cross-task learning, exhibiting significant advantages under high-dimensional feature compression and sparse sample conditions. Although the current model has achieved empirical progress in accuracy and transfer directionality, it still faces limitations in complex action semantic recognition and dynamic feedback adaptation. Future research can further expand in three dimensions: higher-order cross-attention mechanisms, multimodal perception collaborative modeling, and online learning frameworks, driving the evolution of adolescent athletic ability prediction systems toward real-time and intelligent capabilities.

References

1. X. Zheng, V. M. Dwyer, L. A. Barrett, M. Derakhshani, and S. Hu, "Rapid vital sign extraction for real-time opto-physiological monitoring at varying physical activity intensity levels," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3107-3118, 2023. doi: 10.1109/jbhi.2023.3268240
2. M. Molnár, "Sport seasons through the lenses of project management: Identifying the common key characteristics of sport seasons and projects," *Vezetéstudomány-Budapest Management Review*, vol. 55, no. 7-8, pp. 32-44, 2024.
3. N. A. Bhat, and S. U. Farooq, "An empirical evaluation of defect prediction approaches in within-project and cross-project context," *Software Quality Journal*, vol. 31, no. 3, pp. 917-946, 2023. doi: 10.1007/s11219-023-09615-7
4. N. A. Bhat, and S. U. Farooq, "An improved method for training data selection for cross-project defect prediction," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1939-1954, 2022. doi: 10.1007/s13369-021-06088-3
5. S. Tang, S. Huang, and C. Zheng, "A novel cross-project software defect prediction algorithm based on transfer learning," *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 41-57, 2021. doi: 10.26599/tst.2020.9010040
6. K. E. Bennin, A. Tahir, and S. G. MacDonell, "An empirical study on the effectiveness of data resampling approaches for cross-project software defect prediction," *IET Software*, vol. 16, no. 2, pp. 185-199, 2022.
7. S. Panigrahi, N. Stewart, and C. Sripada, "Selective inference for sparse multitask regression with applications in neuroimaging," *The Annals of Applied Statistics*, vol. 18, no. 1, pp. 445-467, 2024. doi: 10.1214/23-aoas1796
8. D. Kodati, and R. Tene, "Advancing mental health detection in texts via multi-task learning with soft-parameter sharing transformers," *Neural Computing and Applications*, vol. 37, no. 5, pp. 3077-3110, 2025. doi: 10.1007/s00521-024-10753-7
9. C. E. Tomlinson, P. J. Laurienti, and R. G. Lyday, "3M_BANTOR: A regression framework for multitask and multisession brain network distance metrics," *Network Neuroscience*, vol. 7, no. 1, pp. 1-21, 2023. doi: 10.1162/netn_a_00274
10. B. Patel, A. Orlichenko, and A. Patel, "Explainable multimodal graph isomorphism network for interpreting sex differences in adolescent neurodevelopment," *Applied Sciences*, vol. 14, no. 10, p. 4144, 2024. doi: 10.3390/app14104144