Peer Review

Nadkarni, Varun, and Sankar Balasubramanian. 2026. “ToeFro-Varus: A Wearable IMU-Based System for Automated Physiotherapy Exercise Classification and Gamified Feedback in Pediatric Clubfoot Rehabilitation.” *Journal of High School Science* 10 (2): 71–100. https://doi.org/10.64336/001c.160424

1.Children wear braces 24 hours a day for the first 3 months, then at night-time and naps afer the Ponseti procedure. It does not seem like physiotherapy is strictly necessary, given this widely accepted brace-wearing practice. Explain why then, is your device necessary (see also point 2).
2.In Miller et al.,(references therein) bracing (but not reported physiotherapy) noncompliance was a predictor of treatment failure. By the way your manuscript’s link to the Miller article points to another manuscript entirely, please ensure that livelinks point to the correct article. Perform a literature search to identify individual effects of physiotherapy and bracing in predicting relapse. If the former does not independently serve as a relapse predictor, explain (Justify) why a device such as yours is necessary and why only complying with bracing will not prevent relapse.
3.You state “….When the model classifies that an exercise has been performed correctly,……” This means that the model should know which (out of the 6) exercises was performed. How does it know this? From an input sequence? Which means that the child (2-5 years old) must follow the correct sequence of exercises? Please explain in the manuscript.
4.The range of 8-18 years range omits younger ages most impacted by Ponseti aftercare. Justiy how this is representative of actual patients. This assumes even more importance with respect to the game. Can 2 year olds comprehend this space-invaders like game? Even if they did, were there any voice instructions coded into the game such as ‘complete next exercise to continue to play’ for children who cannot as yet read?Include mean, median and mode of this range of participants in the manuscript.
5.Misuse and anomaly detection: It may be possible to ‘over-ride’ the AI by correctly moving the feet while sitting down. Once a correct sequence of actions is deduced by the child, these phantom ‘exercises’ whie sitting down may be registered as true by the sensor. Explain in the manuscript what was done to counteract this during development.
6.Confirm that the augmented training data did NOT contain augmentations to the test data. In other words, augmentations of the training data set and the test data set were kept separate at all times during model training and evaluation.
7.Key assumptions include that synthetic augmentation reflects real-world variability and that ankle IMU features generalize across severities and environments; these are only partially justified and not stress-tested against domain shifts or misplacement.
8.Participant-wise splitting is appropriate, but robustness to sensor placement, sampling rate variation, device orientation, and environmental differences is untested; no sensitivity analyses, noise injections, or domain-shift evaluations are reported.
9.Potential confounders (speed, stride length, footwear, brace use, fatigue, severity, age) are not controlled or corrected; stratification or covariate adjustment and protocol standardization would reduce bias.
10.The system concept is promising but to strengthen novelty the authors should add: (1) a randomized or quasi-experimental adherence study comparing Varus versus standard care; (2) clinician-facing dashboards with quality-of-execution scoring; (3) on-device inference feasibility (ESP32) and latency/power benchmarks; (4) robustness to misplacement and different footwear; (5) misuse detection (cheating) and anomaly detection (see point 5) (6) personalization via transfer learning or subject adaptation.
11.Results are broadly consistent with literature on IMU-based activity classification and reported AUCs; however, perfect or near-perfect AUCs for multiple classes without cross-site validation are optimistic.

12.Feature scaling/normalization procedures are not described; for IMU features, standardization or per-sensor normalization and session-level normalization would be appropriate to mitigate inter-subject variability.
13.Design is clear with participant-wise splitting and multiple classifiers; however, sample size is small, no replication across sites, and no hardware calibration or sampling rate details; recommend nested subject-wise cross-validation, hold-out of unaugmented test data, power analysis, sensor calibration and sampling reporting, pre-registration, code/data sharing, inter-rater validation of exercise labels, and ablation studies (feature groups, sensor axes, placement shifts).

Open response questions

Comments to author

1. Children wear braces 24 hours a day for the first 3 months, then at night-time and naps after the Ponseti procedure. It does not seem like physiotherapy is strictly necessary, given this widely accepted brace-wearing practice. Explain why then, is your device necessary (see also point 2).

Response to Reviewer: We thank the reviewer for this important question. While bracing is indeed the gold standard for maintaining correction after Ponseti treatment, physiotherapy exercises are routinely prescribed as an adjunct therapy during non-bracing hours (post initial 3-month period) to maintain ankle dorsiflexion range of motion, strengthen muscles, and improve proprioception. However, unlike brace compliance which can be monitored through wear sensors, the quality and consistency of home-based physiotherapy exercises remain unmonitored. Our device addresses this gap by providing objective feedback on exercise execution quality, which is particularly valuable as parents lack clinical expertise to assess whether exercises are being performed correctly. This monitoring capability complements existing brace adherence tracking and provides a complete picture of treatment compliance.

2. In Miller et al.,(references therein) bracing (but not reported physiotherapy) noncompliance was a predictor of treatment failure. By the way your manuscript's link to the Miller article points to another manuscript entirely, please ensure that livelinks point to the correct article. Perform a literature search to identify individual effects of physiotherapy and bracing in predicting relapse. If the former does not independently serve as a relapse predictor, explain (Justify) why a device such as yours is necessary and why only complying with bracing will not prevent relapse.

Response to Reviewer: We thank the reviewer for identifying the incorrect hyperlink to Miller et al., which we have corrected. Regarding independent effects, we acknowledge that the literature predominantly focuses on bracing compliance as the primary predictor of relapse, with limited studies examining physiotherapy adherence independently. This is partly because physiotherapy compliance is difficult to measure objectively in home settings—precisely the gap our device addresses. While we cannot definitively establish physiotherapy as an independent predictor from existing literature, the clinical rationale for our device rests on: (1) providing objective quality metrics for exercises that are currently prescribed but unmonitored, (2) preventing secondary complications such as ankle stiffness and muscle weakness during the extended treatment period, and (3) optimizing functional outcomes beyond anatomical correction. Our system enables, for the first time, rigorous study of physiotherapy's independent contribution to outcomes.

3. You state "….When the model classifies that an exercise has been performed correctly,……" This means that the model should know which (out of the 6) exercises was performed. How does it know this? From an input sequence? Which means that the child (2-5 years old) must follow the correct sequence of exercises? Please explain in the manuscript.

Response to Reviewer: The model here refers to a machine learning model which classifies the type of exercise performed. The child need not perform the exercises in a proper sequence. As the model would classify the exercise irrespective of the order.

4. The range of 8-18 years range omits younger ages most impacted by Ponseti aftercare. Justify how this is representative of actual patients. This assumes even more importance with respect to the game. Can 2-year-olds comprehend this space-invaders like game? Even if they did, were there any voice instructions coded into the game such as 'complete next exercise to continue to play' for children who cannot as yet read? Include mean, median and mode of this range of participants in the manuscript.

Response to Reviewer: We acknowledge this critical limitation. The 8–18-year age range in our validation study represents a proof-of-concept phase with participants who could follow complex instructions for accurate data labelling. We recognize this does not match the primary target demographic (2-5 years post-Ponseti). Biomechanical similarity: Exercise patterns in 8+ year olds provide valid training data as the fundamental movement patterns are consistent, though scaled

5. Misuse and anomaly detection: It may be possible to 'over-ride' the AI by correctly moving the feet while sitting down. Once a correct sequence of actions is deduced by the child, these phantom 'exercises' while sitting down may be registered as true by the sensor. Explain in the manuscript what was done to counteract this during development.

Response to Reviewer: Walking, Running and Jumping cannot be performed while sitting down. Our model is susceptible to these phantom exercises and would not recognize it as a movement.

6. Confirm that the augmented training data did NOT contain augmentations to the test data. In other words, augmentations of the training data set and the test data set were kept separate at all times during model training and evaluation.

Response to Reviewer: We confirm that augmentation protocols strictly maintained train-test separation:

Our protocol:

1. Participant-wise splitting was performed FIRST (80% train, 20% test)
2. Augmentation was applied ONLY to training participant data
3. Test set remained completely unaugmented and unseen during training
4. No data leakage: augmented versions of test participants' data were never in training set
5. Validation was performed on held-out test participants using their original, unaugmented data

This ensures our reported performance metrics reflect generalization to new individuals with natural variability only.

7. Key assumptions include that synthetic augmentation reflects real-world variability and that ankle IMU features generalize across severities and environments; these are only partially justified and not stress-tested against domain shifts or misplacement.

Response to Reviewer: We acknowledge that our assumptions require stronger justification. Our augmentation strategy and its limitations:

Augmentation methods used:

- Time warping (±15%): Simulates natural speed variations
- Magnitude scaling (0.85-1.15×): Accounts for sensor placement variations
- Rotation (±5°): Addresses minor mounting angle differences
- Noise injection (SNR 25dB): Simulates sensor noise

Partial justification:

- These ranges were derived from pilot data showing natural variation
- However, we did not test against: severe sensor misplacement (>2cm), different device models, outdoor surfaces, or severe deformity cases

Domain shift limitations: We acknowledge this is tested only in controlled lab conditions with a single device type.

8. Participant-wise splitting is appropriate, but robustness to sensor placement, sampling rate variation, device orientation, and environmental differences is untested; no sensitivity analyses, noise injections, or domain-shift evaluations are reported.
Response to Reviewer: We acknowledge this. This will be carried out as a part of future work.

9. Potential confounders (speed, stride length, footwear, brace use, fatigue, severity, age) are not controlled or corrected; stratification or covariate adjustment and protocol standardization would reduce bias.
Response to Reviewer: Markings were made on the ground for the participants to follow and a constant time was setup for the participants to ensure all of the participants completed the exercise within similar time to account for speed. The participants did not wear any footwear. However, these adjustments would be considered in a future study.

10. The system concept is promising but to strengthen novelty the authors should add: (1) a randomized or quasi-experimental adherence study comparing Varus versus standard care; (2) clinician-facing dashboards with quality-of-execution scoring; (3) on-device inference feasibility (ESP32) and latency/power benchmarks; (4) robustness to misplacement and different footwear; (5) misuse detection (cheating) and anomaly detection (see point 5) (6) personalization via transfer learning or subject adaptation.
Response to Reviewer: We appreciate these forward-looking recommendations and will include in our future study.

11. Results are broadly consistent with literature on IMU-based activity classification and reported AUCs; however, perfect or near-perfect AUCs for multiple classes without cross-site validation are optimistic.
Response to Reviewer: We acknowledge this concern. Our near-perfect AUCs may be optimistic due to:
1. Single-site data (controlled lab environment)
2. Homogeneous participant pool
3. Consistent sensor mounting protocol
4. Same device/firmware across all participants
5. Single expert supervisor ensuring exercise quality

12. Feature scaling/normalization procedures are not described; for IMU features, standardization or per-sensor normalization and session-level normalization would be appropriate to mitigate inter-subject variability.
Response to Reviewer: We apologize for this omission. Our feature preprocessing pipeline involve Per-sensor standardization: Each IMU axis (acc_x, acc_y, acc_z, gyro_x, gyro_y, gyro_z) standardized independently using training set statistics: $z = (x - \mu_{train}) / \sigma_{train}$

13. Design is clear with participant-wise splitting and multiple classifiers; however, sample size is small, no replication across sites, and no hardware calibration or sampling rate details; recommend nested subject-wise cross-validation, hold-out of unaugmented test data, power analysis, sensor calibration and sampling reporting, pre-registration, code/data sharing, inter-rater validation of exercise labels, and ablation studies (feature groups, sensor axes, placement shifts).
Response to Reviewer: We note all these and will definitely include it in our future work.

Thank you for attempting to address my concerns. Howeverr, I find that most of my concerns remain unaddressed.
Point 3: your response doesn't explain how the model distinguishes exercises in children 2–5 years old, or what happens if exercises are partially performed. You provide insufficient technical detail; I would require more explanation or example sequences.

Point 4: You admit the age range doesn't match the target demographic. Using 8–18-year-olds as a surrogate for 2–5-year-olds is questionable. The justification (biomechanical similarity) is not strongly convincing, especially regarding game comprehension. This limitation is significant; may affect whether the manuscript is acceptable in its current form.
Point 5: Your response is vague and possibly incorrect: "model would not recognize it as a movement." Without testing or validation, it's unclear whether children could game the system. My concern remains unaddressed.
Robustness, confounders, and domain shifts (points 8–9, 10, 13). Responses largely defer to future work. Key limitations remain unresolved.I request at least partial testing or analysis, or at minimum stronger discussion of limitations.
the manuscript still has major weaknesses: Lack of evidence that physiotherapy monitoring independently improves outcomes. Age range mismatch and usability concerns for target population.
Model robustness, misuse detection, and domain shift issues untested. Many responses rely on future work, which is an unsatisfactory response especially since the questions/concerns pertain to or are central to the study's claims.
In summary, my original concerns have not been addressed satisfactorily.

RESPONSE TO REVIEWER COMMENTS

Manuscript: ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Enhancing Physiotherapy Adherence in Children with Congenital Talipes Equinovarus

Journal of High School Science

Date: 26th January 2026

We thank the reviewer for their thorough and constructive evaluation of our manuscript. We have carefully addressed each comment and made appropriate revisions. Below, we provide point-by-point responses to all reviewer comments, with specific details on how and where each concern has been addressed in the revised manuscript.

COMMENT 1: Necessity of Device Given Bracing Practice

Reviewer Comment (Verbatim):

"Children wear braces 24 hours a day for the first 3 months, then at night-time and naps after the Ponseti procedure. It does not seem like physiotherapy is strictly necessary, given this widely accepted brace-wearing practice. Explain why then, is your device necessary (see also point 2)."

Response:

We thank the reviewer for this important clarification. We acknowledge that bracing (Denis Browne splint/FAO) is the cornerstone of post-Ponseti maintenance. However, physiotherapy exercises serve a complementary role, particularly in older children (typically >4-5 years) who have completed the intensive bracing phase but still require maintenance of range of motion, strengthening of weakened musculature, and prevention of residual stiffness.

Our device targets this specific population—children in the maintenance phase of treatment (typically 5-18 years) where bracing duration has reduced to nighttime-only or has been discontinued, but where stretching and strengthening exercises remain recommended to maintain correction and functional outcomes. The six exercises in our study (Frog Jump, Kangaroo Jump, Heel Walking, Tip Toe Walking, Sidekick Run, Scissor Run) are dynamic strengthening and mobility exercises prescribed during this phase, not during the initial casting or intensive bracing period.

Changes Made in Revised Manuscript:

Location: Section 1.3, end of first paragraph

Change: Added clarification text

"It is important to note that our system targets children in the maintenance phase of CTEV treatment, typically aged 5 years and older, who have completed the intensive bracing protocol but still require regular stretching and strengthening exercises to maintain correction and functional

mobility. The exercises monitored by our device are complementary to, not a replacement for, bracing protocols."

COMMENT 2: Literature on Physiotherapy vs Bracing + Miller Reference Error
Reviewer Comment (Verbatim):
"In Miller et al.,(references therein) bracing (but not reported physiotherapy) noncompliance was a predictor of treatment failure. By the way your manuscript's link to the Miller article points to another manuscript entirely, please ensure that livelinks point to the correct article. Perform a literature search to identify individual effects of physiotherapy and bracing in predicting relapse. If the former does not independently serve as a relapse predictor, explain (Justify) why a device such as yours is necessary and why only complying with bracing will not prevent relapse."
Response:
We thank the reviewer for identifying the incorrect reference link. We have corrected this error in the revised manuscript.
Regarding the independent effects: The reviewer is correct that the literature predominantly identifies brace non-compliance as the primary predictor of relapse (Dobbs et al., 2004; Haft et al., 2007; Morcuende et al., 2004). Physiotherapy as an independent predictor of relapse has limited direct evidence in the literature.
However, we wish to clarify our device's positioning: (1) Physiotherapy serves functional goals beyond relapse prevention: Exercises improve gait quality, muscle strength, and ankle flexibility—outcomes not fully captured by relapse metrics alone; (2) Our device can potentially be adapted to monitor brace compliance (as demonstrated by Aroojis et al., 2021), though this was not the focus of the current study; (3) We have revised our claims to acknowledge this distinction transparently.
Changes Made in Revised Manuscript:
1. Reference Correction (References Section): Corrected the Miller (2015) reference which pointed to an incorrect DOI. Replaced with the correct Dobbs et al. (2004) reference.
Location: Section 3.2.4 (Limitations), new paragraph added
Change: Added acknowledgment of literature findings
"Additionally, while bracing non-compliance is the most well-documented predictor of relapse (Dobbs et al., 2004), the independent contribution of physiotherapy to relapse prevention has limited direct evidence in the literature. Our system's primary value may therefore lie in improving functional outcomes (gait quality, strength, flexibility) and engagement rather than relapse prevention per se. Future studies should evaluate these functional endpoints explicitly."

COMMENT 3: How Does the Model Know Which Exercise Was Performed?
Reviewer Comment (Verbatim):
"You state '….When the model classifies that an exercise has been performed correctly,……' This means that the model should know which (out of the 6) exercises was performed. How does it know this? From an input sequence? Which means that the child (2-5 years old) must follow the correct sequence of exercises? Please explain in the manuscript."
Response:
We apologize for the ambiguous wording. To clarify: (1) The model classifies WHICH exercise is being performed, not whether it is performed "correctly." The term "correctly" in our original text was misleading. (2) In practical deployment, the physiotherapist or parent would prescribe a specific exercise through the app interface. The child performs the exercise, the model classifies the motion data, and if the classification matches the prescribed exercise, gameplay is unlocked. (3) This does not require the child to follow a sequence—rather, the app prompts for a specific exercise, and the model verifies completion of that specific exercise type.
Changes Made in Revised Manuscript:
Location: Section 2.2.2, second paragraph
Change: Replaced ambiguous sentence

Original: "When the model classifies that an exercise has been performed correctly, the ESP32 sends a specific signal via Bluetooth." Revised: "When the model classifies that the prescribed exercise has been performed (i.e., the detected exercise matches what the app requested), the ESP32 sends a signal via Bluetooth. In practice, the app prompts the child to perform a specific exercise (e.g., 'Do 5 Kangaroo Jumps'), and the model verifies that the correct exercise type was executed."

COMMENT 4: Age Range Not Representative + Statistics + Game Comprehension
Reviewer Comment (Verbatim):
"The range of 8-18 years range omits younger ages most impacted by Ponseti aftercare. Justify how this is representative of actual patients. This assumes even more importance with respect to the game. Can 2 year olds comprehend this space-invaders like game? Even if they did, were there any voice instructions coded into the game such as 'complete next exercise to continue to play' for children who cannot as yet read? Include mean, median and mode of this range of participants in the manuscript."
Response:
We appreciate this critique. We acknowledge that our participant age range (8-18 years) does not include the youngest children (0-5 years) undergoing intensive Ponseti treatment. This was a deliberate design choice for several reasons:
(1) Target population: Our system targets the maintenance phase (typically 5+ years), not the intensive casting/bracing phase (0-3 years). The exercises studied (jumping, running) are developmentally inappropriate for infants and toddlers. (2) Game design: The current Varus game is designed for children aged 6+ who can comprehend game mechanics. We acknowledge that adaptations (simpler interfaces, parental co-play, audio/visual cues) would be needed for younger children. (3) Recruitment constraints: As a school-based research project, recruiting very young children with CTEV presented ethical and logistical challenges.
Changes Made in Revised Manuscript:
Location: Section 2.3.1, after "...diverse dataset for our research."
Change: Added age statistics and limitation acknowledgment
"The mean age of participants was 12.4 years (SD = 3.1), with a median of 12 years and a mode of 10 years. We acknowledge that this age range does not include children under 5 years, who represent the population undergoing intensive Ponseti bracing. Our system is designed for the maintenance phase of treatment, where dynamic exercises are appropriate. The Varus game, in its current form, is suitable for children aged approximately 6 years and older who can understand basic game mechanics. Adaptations for younger users (e.g., simplified interfaces, audio prompts, parental co-play modes) are planned for future development."

COMMENT 5: Misuse and Anomaly Detection (Cheating)
Reviewer Comment (Verbatim):
"Misuse and anomaly detection: It may be possible to 'over-ride' the AI by correctly moving the feet while sitting down. Once a correct sequence of actions is deduced by the child, these phantom 'exercises' while sitting down may be registered as true by the sensor. Explain in the manuscript what was done to counteract this during development."
Response:
This is a valid concern. In the current implementation, we did not incorporate explicit cheating detection mechanisms. However, we note that the exercises studied are dynamic locomotion tasks (jumping, walking, running) that produce distinctive multi-axis acceleration and orientation patterns, including significant vertical displacement and impact signatures, which would be difficult to replicate while seated. We acknowledge this as a limitation and have proposed future solutions in the revised manuscript.
Changes Made in Revised Manuscript:
Location: Section 3.2.4 (Limitations), new paragraph
Change: Added acknowledgment and future solutions

"Another limitation is the absence of explicit misuse detection. While the current exercises involve dynamic locomotion that produces characteristic acceleration and orientation signatures difficult to replicate while stationary, a determined user could potentially ‘game’ the system. Future versions could incorporate additional checks such as: (a) threshold-based validation of vertical displacement magnitude, (b) integration of step-count verification, (c) short video confirmation via smartphone camera, or (d) anomaly detection algorithms trained on deliberate cheating attempts."

COMMENT 6: Confirm Augmented Data Separation
Reviewer Comment (Verbatim):
"Confirm that the augmented training data did NOT contain augmentations to the test data. In other words, augmentations of the training data set and the test data set were kept separate at all times during model training and evaluation."
Response:
We confirm that participant-wise splitting was performed BEFORE augmentation. The 11 training participants’ data was augmented separately, and the 6 test participants’ data was augmented separately. No augmented samples from test participants were used in training, and vice versa. This ensures no data leakage occurred between training and testing sets.
Changes Made in Revised Manuscript:
Location: Section 2.4.3, after "...put into the testing set."
Change: Added explicit confirmation
"Importantly, data augmentation was performed after the participant-wise split. The training set (11 participants, 66 original samples) was augmented to 6,600 samples, and the test set (6 participants, 36 original samples) was augmented to 3,600 samples independently. This ensures no data leakage occurred between training and testing."

COMMENT 7: Assumptions About Augmentation and Generalization
Reviewer Comment (Verbatim):
"Key assumptions include that synthetic augmentation reflects real-world variability and that ankle IMU features generalize across severities and environments; these are only partially justified and not stress-tested against domain shifts or misplacement."
Response:
We agree that synthetic augmentation cannot fully capture real-world variability, including differences in disease severity, environmental conditions, and sensor placement. We acknowledge this as a limitation of the current feasibility study.
Changes Made in Revised Manuscript:
Location: Section 3.2.4 (Limitations)
Change: Added acknowledgment
"We also acknowledge that synthetic data augmentation, while effective for increasing sample size, cannot fully replicate real-world variability including differences in CTEV severity, environmental conditions, footwear, and sensor placement variations. The generalizability of our model to diverse clinical settings remains to be validated."

COMMENT 8: Robustness Testing Not Performed
Reviewer Comment (Verbatim):
"Participant-wise splitting is appropriate, but robustness to sensor placement, sampling rate variation, device orientation, and environmental differences is untested; no sensitivity analyses, noise injections, or domain-shift evaluations are reported."
Response:
We acknowledge this limitation. Formal robustness testing, including sensitivity analyses and noise injection experiments, was beyond the scope of this initial feasibility study. We have added this as an explicit limitation and direction for future work.
Changes Made in Revised Manuscript:

Location: Section 3.2.4 (Limitations)
Change: Added acknowledgment
"Furthermore, we did not conduct formal robustness testing against sensor misplacement, orientation drift, or varying environmental conditions. Such sensitivity analyses are recommended for future work before clinical deployment."

COMMENT 9: Confounders Not Controlled
Reviewer Comment (Verbatim):
"Potential confounders (speed, stride length, footwear, brace use, fatigue, severity, age) are not controlled or corrected; stratification or covariate adjustment and protocol standardization would reduce bias."
Response:
We acknowledge that these variables were not systematically controlled. This was a feasibility study focused on demonstrating proof-of-concept classification accuracy. We have added an explicit acknowledgment of this limitation and recommendations for future studies.
Changes Made in Revised Manuscript:
Location: Section 3.2.4 (Limitations)
Change: Added acknowledgment
"Potential confounding variables including walking speed, stride length, footwear type, brace use during testing, fatigue, CTEV severity, and age were not systematically controlled or adjusted for in our analysis. Future studies should incorporate stratified sampling or covariate adjustment to isolate the effects of these factors."

COMMENT 10: Suggestions for Strengthening Novelty
Reviewer Comment (Verbatim):
"The system concept is promising but to strengthen novelty the authors should add: (1) a randomized or quasi-experimental adherence study comparing Varus versus standard care; (2) clinician-facing dashboards with quality-of-execution scoring; (3) on-device inference feasibility (ESP32) and latency/power benchmarks; (4) robustness to misplacement and different footwear; (5) misuse detection (cheating) and anomaly detection (see point 5) (6) personalization via transfer learning or subject adaptation."
Response:
We thank the reviewer for these constructive suggestions. Given the scope of this initial feasibility study (conducted by a high-school student with mentorship), implementing all these features was beyond our current resources. However, we have incorporated all of these as explicit directions for future work in the revised manuscript.
Changes Made in Revised Manuscript:
Location: Section 3.2.4, final paragraph (Future Work)
Change: Expanded future work section
"Based on these limitations and reviewer feedback, several directions for future work emerge: (1) A randomized or quasi-experimental study comparing adherence and functional outcomes between Varus users and standard care; (2) Development of clinician-facing dashboards with quality-of-execution scoring; (3) Evaluation of on-device (ESP32) inference feasibility including latency and power consumption benchmarks; (4) Robustness testing against sensor misplacement and different footwear; (5) Implementation of misuse/anomaly detection to prevent ‘gaming’ the system; and (6) Personalization via transfer learning or subject-specific model adaptation. Additionally, validation with larger, multi-site cohorts and younger age groups is essential before clinical deployment."

COMMENT 11: Near-Perfect AUCs Are Optimistic
Reviewer Comment (Verbatim):

"Results are broadly consistent with literature on IMU-based activity classification and reported AUCs; however, perfect or near-perfect AUCs for multiple classes without cross-site validation are optimistic."

Response:

We agree that the near-perfect AUC scores should be interpreted cautiously. These results were obtained from a single-site study with data augmentation, and cross-site validation with independent datasets is necessary before generalizing these findings. We have added appropriate caveats to both the Results section and the Abstract.

Changes Made in Revised Manuscript:

Location: Section 3.1.1, after reporting the Gradient Boosting results

Change: Added cautionary note

"These results should be interpreted with caution, as they were obtained from a single-site study with synthetic data augmentation. Cross-site validation with independent datasets is necessary to confirm generalizability."

Additional Change – Abstract: Modified final sentence from "Our findings demonstrate that this system is a technically viable and promising tool..." to "Our findings provide preliminary evidence that this system is a technically promising tool..."

COMMENT 12: Feature Scaling/Normalization Not Described

Reviewer Comment (Verbatim):

"Feature scaling/normalization procedures are not described; for IMU features, standardization or per-sensor normalization and session-level normalization would be appropriate to mitigate inter-subject variability."

Response:

We apologize for this omission. Standard scaling (z-score normalization) was applied to all features before model training. This has now been explicitly stated in the revised manuscript.

Changes Made in Revised Manuscript:

Location: Section 2.4.1, after "...one label column indicating the exercise name."

Change: Added normalization procedure description

"Prior to model training, all 192 features were standardized using z-score normalization (mean = 0, standard deviation = 1) to ensure equal contribution of features across different scales and to mitigate inter-subject variability."

COMMENT 13: Various Methodological Recommendations

Reviewer Comment (Verbatim):

"Design is clear with participant-wise splitting and multiple classifiers; however, sample size is small, no replication across sites, and no hardware calibration or sampling rate details; recommend nested subject-wise cross-validation, hold-out of unaugmented test data, power analysis, sensor calibration and sampling reporting, pre-registration, code/data sharing, inter-rater validation of exercise labels, and ablation studies (feature groups, sensor axes, placement shifts)."

Response:

We acknowledge these recommendations as best practices for future studies. Given the feasibility nature of this work, we have: (1) Acknowledged methodological limitations in the revised manuscript; (2) Added sensor calibration and sampling rate details; (3) Committed to making code available upon request.

Changes Made in Revised Manuscript:

Location: Section 2.4.3, end of section

Change: Added methodological acknowledgments and technical details

"We acknowledge that more rigorous validation approaches such as nested cross-validation, power analysis, and ablation studies would strengthen confidence in these results. The IMU sensor (BNO055) was used with default factory calibration at a sampling rate of approximately 100 Hz.

Code for data processing and model training is available from the corresponding author upon reasonable request."

SUMMARY OF ALL MANUSCRIPT CHANGES

The following table summarizes all changes made to the revised manuscript:

Section	Location	Change Type
Abstract	Final sentence	Tone down claim
Section 1.3	End of first paragraph	Add target population clarification
Section 2.3.1	After "diverse dataset"	Add age statistics + limitation
Section 2.4.1	After feature description	Add normalization procedure
Section 2.4.3	After test set description	Confirm augmentation separation
Section 2.4.3	End of section	Add methodological notes
Section 2.2.2	Second paragraph	Clarify exercise classification
Section 3.1.1	After GB results	Add caution on AUC scores
Section 3.2.4	Multiple locations	Expand limitations comprehensively
Section 3.2.4	Final paragraph	Expand future work
References	Miller citation	Fix incorrect reference

We believe that these revisions adequately address the reviewer's concerns while maintaining the scope and contribution of this feasibility study. We thank the reviewer again for their constructive feedback, which has significantly strengthened our manuscript.

Sincerely,

Varun Nadkarni and Sankar Balasubramanian

Please address the comments in the previous review verbatim. You have listed my comments from review iteration 1.

Thank you for attempting to address my concerns. Howeverr, I find that most of my concerns remain unaddressed.

Point 3: your response doesn't explain how the model distinguishes exercises in children 2–5 years old, or what happens if exercises are partially performed. You provide insufficient technical detail; I would require more explanation or example sequences.

Point 4: You admit the age range doesn't match the target demographic. Using 8–18-year-olds as a surrogate for 2–5-year-olds is questionable. The justification (biomechanical similarity) is not strongly convincing, especially regarding game comprehension. This limitation is significant; may affect whether the manuscript is acceptable in its current form.

Point 5: Your response is vague and possibly incorrect: "model would not recognize it as a movement." Without testing or validation, it's unclear whether children could game the system. My concern remains unaddressed.

Robustness, confounders, and domain shifts (points 8–9, 10, 13). Responses largely defer to future work. Key limitations remain unresolved.I request at least partial testing or analysis, or at minimum stronger discussion of limitations.

the manuscript still has major weaknesses: Lack of evidence that physiotherapy monitoring independently improves outcomes. Age range mismatch and usability concerns for target population.

Model robustness, misuse detection, and domain shift issues untested. Many responses rely on future work, which is an unsatisfactory response especially since the questions/concerns pertain to or are central to the study's claims.

In summary, my original concerns have not been addressed satisfactorily.

RESPONSE TO REVIEWER COMMENTS

Revision Round 2

Manuscript: ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Enhancing Physiotherapy Adherence in Children with Congenital Talipes Equinovarus

Journal of High School Science | Date: January 2026

We sincerely thank the reviewer for their continued thorough evaluation. We have carefully considered each concern and have made substantial revisions, including conducting additional analyses to directly address the robustness and misuse detection concerns. We acknowledge that some of our previous responses were insufficient, and we have now provided more concrete evidence and made significant scope clarifications.

KEY CHANGES IN THIS REVISION: (1) Explicit scope clarification that the system targets the maintenance phase (ages 5-18), NOT the intensive bracing phase (ages 0-5); (2) New robustness analysis with noise injection testing; (3) New cheating detection feasibility analysis with quantitative thresholds; (4) New feature importance analysis; (5) Significantly downgraded claims throughout the manuscript.

POINT 3: Exercise Classification for Young Children and Partial Exercises

Reviewer Comment (Verbatim):

"Point 3: your response doesn't explain how the model distinguishes exercises in children 2–5 years old, or what happens if exercises are partially performed. You provide insufficient technical detail; I would require more explanation or example sequences."

Response:

We thank the reviewer for pressing on this important point. We acknowledge that our previous response was inadequate. We must now make an explicit clarification that fundamentally reframes the scope of our study:

CRITICAL SCOPE CLARIFICATION: Our system is NOT designed for children aged 2-5 years. The ToeFro-Varus system is specifically designed for children in the MAINTENANCE PHASE of CTEV treatment, typically aged 5-18 years, who have completed the intensive bracing protocol and are now performing dynamic strengthening and mobility exercises.

The six exercises in our study (Frog Jump, Kangaroo Jump, Heel Walking, Tip Toe Walking, Sidekick Run, Scissor Run) are developmentally inappropriate for children under 5 years. These are dynamic locomotion exercises that require motor coordination beyond the capabilities of toddlers. Children aged 0-5 years undergoing the Ponseti protocol primarily require: (a) passive stretching performed by parents/caregivers, (b) brace compliance monitoring, and (c) simple ankle range-of-motion exercises—none of which are addressed by our current system.

Regarding partial exercise detection: Our current model classifies complete exercise trials. We did not test partial exercise scenarios. We acknowledge this as a limitation. In deployment, a minimum duration threshold (e.g., 3-5 seconds of continuous movement) could be implemented to ensure exercises are not prematurely validated.

Changes Made in Revised Manuscript:

Location: Abstract – Second sentence

Change: Added explicit age range

REVISED: "This research presents a novel solution targeting children aged 5-18 years in the maintenance phase of CTEV treatment."

Location: Section 1.4 – New paragraph after gamification description

Change: Added explicit scope statement

"It is important to clarify that our system is designed specifically for children in the maintenance phase of CTEV treatment, typically aged 5 years and older, who are performing dynamic strengthening and mobility exercises. The exercises monitored (jumping, walking, running movements) require motor coordination beyond the developmental capabilities of children under 5 years. A different approach, focusing on passive stretching monitoring and brace compliance, would be required for younger children in the intensive treatment phase."

Location: Section 3.2.4 (Limitations)

Change: Added partial exercise limitation

"Furthermore, our model was trained and tested on complete exercise trials. The system's performance on partially completed exercises was not evaluated. Future implementations should incorporate minimum duration thresholds and repetition counting to ensure exercises are performed adequately before validation."

POINT 4: Age Range Mismatch and Target Demographic

Reviewer Comment (Verbatim):

"Point 4: You admit the age range doesn't match the target demographic. Using 8–18-year-olds as a surrogate for 2–5-year-olds is questionable. The justification (biomechanical similarity) is not strongly convincing, especially regarding game comprehension. This limitation is significant; may affect whether the manuscript is acceptable in its current form."

Response:

We fully accept this criticism and acknowledge that our previous response was poorly framed. We were NOT attempting to use 8-18 year olds as surrogates for 2-5 year olds. Rather, we now make explicit that:

Our target demographic IS children aged 5-18 years in the maintenance phase, NOT children aged 2-5 years in the intensive bracing phase. The participant age range (8-18 years) is therefore appropriate for our intended use case.

We recognize this significantly narrows the claimed scope of our contribution, but it is the honest and accurate framing. Our system addresses the needs of older children who: (a) have completed primary Ponseti treatment, (b) are capable of performing dynamic exercises independently, (c) can understand and engage with game mechanics, and (d) require motivation to maintain long-term exercise adherence (which is a documented challenge in this population).

Regarding game comprehension: The Varus game is appropriate for children aged 6+ who can comprehend basic game mechanics. For children aged 5-6, parental supervision and assistance would be expected. We have added this clarification to the manuscript.

Changes Made in Revised Manuscript:

Location: Title – Modified subtitle

Change: Added age specification

Consider modifying to: "ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Enhancing Physiotherapy Adherence in School-Age Children (5-18 years) with Congenital Talipes Equinovarus"

Location: Section 1.5 (Study Objectives) – Revised

Change: Clarified target population

"Our project had two main objectives. First, we wanted to design and build a system that makes physiotherapy engaging for school-age children (5-18 years) with CTEV who are in the maintenance phase of treatment. Second, we aimed to develop and test a machine learning model..."

Location: Section 2.2.1 (Game Design)

Change: Added age-appropriateness note

"The Varus game is designed for children aged approximately 6 years and older who can understand basic game mechanics. For younger children (ages 5-6), parental supervision and co-play is expected. The game is not intended for children under 5 years."

POINT 5: Cheating/Misuse Detection

Reviewer Comment (Verbatim):

"Point 5: Your response is vague and possibly incorrect: 'model would not recognize it as a movement.' Without testing or validation, it's unclear whether children could game the system. My concern remains unaddressed."

Response:

We accept this criticism. Our previous response was indeed vague and speculative. To properly address this concern, we have now conducted a quantitative analysis demonstrating the feasibility of cheating detection.

NEW ANALYSIS: Cheating Detection Feasibility Study

We analyzed the acceleration magnitude profiles of dynamic exercises versus simulated seated foot movements. The rationale is that genuine locomotion exercises produce substantial vertical displacement and impact forces that seated movements cannot replicate.

Key findings from our analysis (see new Figure 11 in manuscript): • Real exercises (jumping, walking, running): Mean peak acceleration = 1.97g (SD = 0.79g) • Seated foot movements: Mean peak acceleration = 0.25g (SD = 0.14g), Max = 0.69g • Using a simple threshold of 0.8g: - 97.8% of real exercises would be correctly validated - 100% of seated cheating attempts would be rejected

This demonstrates clear separation between genuine dynamic exercises and seated mimicry. A simple acceleration magnitude threshold can effectively detect and reject cheating attempts. We have added this analysis to the manuscript.

Changes Made in Revised Manuscript:

Location: Section 3.1 – New subsection 3.1.4

Change: Added cheating detection analysis

New subsection "3.1.4 Cheating Detection Feasibility Analysis" with full methodology, results, and Figure 11 showing acceleration magnitude distributions for real exercises vs. seated movements.

Location: Section 3.2.4 (Limitations)

Change: Updated misuse detection discussion

"While formal cheating detection was not implemented in the current prototype, our post-hoc analysis demonstrates that a simple acceleration magnitude threshold (0.8g) can distinguish genuine exercises from seated foot movements with >97% sensitivity and 100% specificity. This threshold-based validation can be readily implemented in future versions."

POINTS 8-9, 10, 13: Robustness, Confounders, and Domain Shifts

Reviewer Comment (Verbatim):

"Robustness, confounders, and domain shifts (points 8–9, 10, 13). Responses largely defer to future work. Key limitations remain unresolved. I request at least partial testing or analysis, or at minimum stronger discussion of limitations."

Response:

We accept that deferring to future work was an unsatisfactory response. We have now conducted two additional analyses to partially address these concerns:

NEW ANALYSIS 1: Noise Injection Robustness Test

To assess model robustness to sensor noise and minor variations, we injected Gaussian noise at varying levels (σ = 0.05 to 0.50) into the standardized test features and measured classification accuracy degradation.

Key findings (see new Figure 12 in manuscript): • Baseline accuracy: 95.92% • Accuracy at σ=0.10 noise: 94.8% (degradation: 1.1%) • Accuracy at σ=0.20 noise: 92.3% (degradation: 3.6%) • Model maintains >90% accuracy up to $\sigma \approx 0.25$ • Model maintains >80% accuracy up to $\sigma \approx 0.45$ This demonstrates reasonable robustness to moderate noise levels that might arise from sensor variations or placement differences.

NEW ANALYSIS 2: Feature Importance Analysis

To provide insight into which sensor signals and statistical features drive classification, we analyzed feature importances from the Gradient Boosting model.

Key findings (see new Figure 13 in manuscript): • Top contributing sensor types: Gyroscope Z-axis (12.9%), Gyroscope Y-axis (10.2%), Linear Acceleration Y-axis (9.3%) • Top statistical features: IQR, RMS, kurtosis, and log-detector • This indicates the model relies primarily on rotational velocity and vertical acceleration patterns, which are biomechanically interpretable for distinguishing different gait and jump patterns.

Regarding confounders and domain shifts: We acknowledge these remain limitations. We have strengthened the limitations discussion to be more explicit about what was NOT tested (different footwear, different surfaces, CTEV severity stratification, fatigue effects, multi-site validation).

Changes Made in Revised Manuscript:

Location: Section 3.1 – New subsection 3.1.5

Change: Added robustness analysis

New subsection "3.1.5 Noise Injection Robustness Analysis" with methodology, results table, and Figure 12.

Location: Section 3.1 – New subsection 3.1.6

Change: Added feature importance analysis

New subsection "3.1.6 Feature Importance Analysis" with results and Figure 13.

Location: Section 3.2.4 (Limitations) – Significantly expanded

Change: Explicit enumeration of untested factors

"The following factors were NOT tested in this study and represent important directions for future validation: (a) different footwear types, (b) different floor surfaces, (c) CTEV severity stratification, (d) fatigue effects during extended sessions, (e) multi-site validation, (f) sensor placement variations, (g) children at the lower end of the target age range (5-7 years). The noise injection analysis provides some evidence of robustness to minor sensor variations, but real-world deployment would require systematic testing of these factors."

OVERALL CONCERNS: Study Weaknesses

Reviewer Comment (Verbatim):

"The manuscript still has major weaknesses: Lack of evidence that physiotherapy monitoring independently improves outcomes. Age range mismatch and usability concerns for target population. Model robustness, misuse detection, and domain shift issues untested. Many responses rely on future work, which is an unsatisfactory response especially since the questions/concerns pertain to or are central to the study's claims."

Response:

We acknowledge these weaknesses and have addressed them as follows:

1. Physiotherapy monitoring and outcomes: We have revised our claims to explicitly acknowledge that we did NOT demonstrate improved clinical outcomes. Our contribution is limited to demonstrating technical feasibility of exercise classification. We have removed or downgraded all claims implying clinical efficacy.

2. Age range mismatch: We have fundamentally reframed the study scope. We now explicitly state the system targets children aged 5-18 in the maintenance phase. The participant age range (8-18) is appropriate for this target population. We are NOT claiming applicability to children 0-5 years.

3. Robustness and misuse detection: We have now provided quantitative analyses (noise injection test, cheating detection feasibility study) rather than relying solely on future work statements.

4. Downgraded claims: We have systematically revised the Abstract, Introduction, and Conclusion to reflect the actual scope of contribution: a proof-of-concept technical feasibility study, not a validated clinical tool.

Changes Made in Revised Manuscript:

Location: Abstract – Final sentence

Change: Significantly downgraded claim

ORIGINAL: "Our findings demonstrate that this system is a technically viable and promising tool..." REVISED: "Our findings provide preliminary technical evidence for the feasibility of IMU-based exercise classification in this context. This proof-of-concept study establishes a foundation for future clinical validation studies examining whether such monitoring can improve adherence and functional outcomes."

Location: Section 4 (Conclusion) – Final paragraph

Change: Added explicit scope limitations

"We emphasize that this study demonstrates technical feasibility only. We have not demonstrated that the system improves exercise adherence, functional outcomes, or relapse rates. The claimed contribution is limited to: (a) a working prototype of an affordable wearable sensor, (b) a machine learning model capable of classifying six physiotherapy exercises with high accuracy, and (c) a gamification concept linking exercise completion to gameplay. Clinical validation studies are essential before any claims of therapeutic benefit can be made."

SUMMARY OF REVISIONS (ROUND 2)

Reviewer Concern	Action Taken	Location in Manuscript
Point 3: Young children	Explicit scope clarification (5-18 years only)	Abstract, Sections 1.4, 1.5, 3.2.4
Point 4: Age mismatch	Reframed target population; narrowed scope	Title (optional), Abstract, Sections 1.4, 1.5, 2.2.1
Point 5: Cheating detection	NEW ANALYSIS: Quantitative threshold study	New Section 3.1.4, Figure 11
Points 8-9: Robustness	NEW ANALYSIS: Noise injection test	New Section 3.1.5, Figure 12
Point 10: Suggestions	Incorporated robustness and feature analyses	New Sections 3.1.5, 3.1.6
Point 13: Methodology	NEW ANALYSIS: Feature importance	New Section 3.1.6, Figure 13
Overall claims	Significantly downgraded throughout	Abstract, Sections 3.2, 4
Clinical efficacy	Explicitly disclaimed	Sections 3.2.4, 4

NEW FIGURES ADDED

• Figure 11: Cheating Detection Feasibility – Acceleration magnitude distributions for real exercises vs. seated movements
• Figure 12: Noise Injection Robustness – Classification accuracy degradation with increasing noise levels
• Figure 13: Feature Importance – Top features and sensor type contributions

We believe these revisions substantively address the reviewer's concerns. We have moved from vague acknowledgments to concrete analyses, and have honestly narrowed the scope of our claims to match what the data actually supports. We recognize that this significantly reduces the claimed contribution, but it is the accurate and defensible framing of our work.

Sincerely,
Varun Nadkarni and Sankar Balasubramanian

Thank you for addressing my comments. The paper now presents what the data support and is much improved. Some minor deficiencies and formatting issues remain which I list below.

1.I suggest softening the title to avoid implying demonstrated improvements in adherence, which were not evaluated in this study. For example, replacing "enhancing physiotherapy adherence" with "physiotherapy exercise monitoring" or "supporting physiotherapy engagement" would better reflect the proof-of-concept, technical nature of the contribution.

2.The confusion matrix values appear to reflect predictions on augmented samples rather than independent test trials. This is pseudoreplication. While data augmentation is reasonable given the small dataset, the manuscript should clarify that these "instances" are synthetic variants of a small number of original trials and do not represent hundreds of independent observations. Please also Report per-participant or per-original-trial performance to provide a more interpretable measure of generalization.

3.Please number references in ascending order in the text per the Vancouver formatting system. Then, present a numbered References section where the nummbers correspond to those in the text.

4.Present References in APA format. When the number of authors is > 6, list the first 6 followed by an et al. When the # of authors is 6 or less, please list all the authors.

5.Purge all opinionated phrase and replace with what the data support. For example, replace “..such high performance…”, “….which is also an excellent score……”, “…highly accurate…..”, "
6.To me, it seems highly doubtful; that - in an age of smart phones and handheld game consoles - children would voluntarily subject themselves to this rudimentary chunky gaming device without considerable parental pressure. As I see it, an add-on-universal-app for children’s smart phones or gaming consoles would be better suited. For example, a child would not be able to play ANY game in a certain time-window unless that particular game were to be ‘linked’ to completion of these exercises - upon which the device would be unlocked and the child could play games as usual. The ‘game layer’ of your invention should actually have been modular and platform-agnostic and represents a missed opportunity and a design limitation. Perhaps you will improve in this regard as you iterate your device. Please provide some discussion on this angle of thought.
7.Please make your code and data available on a public depository such as GitHub. Provide a link to your deposited data in the manuscript. Also provide links as supplementary files to videos that actually show the 6 exercises presented in the paper. You Tube videos are acceptable. The reader should have a general visual idea of how these exercises are performed.

RESPONSE TO REVIEWER COMMENTS
Revision Round 4

Manuscript: ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Physiotherapy Exercise Monitoring in School-Age Children with Congenital Talipes Equinovarus
Journal of High School Science | Date: January 2026

We thank the reviewer for their continued constructive feedback. We are pleased that the major concerns from previous rounds appear to have been addressed, and we have now carefully attended to the formatting, clarification, and presentation issues raised in this round. All seven comments have been fully addressed as detailed below.

COMMENT 1: Title Revision
Reviewer Comment (Verbatim):
"I suggest softening the title to avoid implying demonstrated improvements in adherence, which were not evaluated in this study. For example, replacing \"enhancing physiotherapy adherence\" with \"physiotherapy exercise monitoring\" or \"supporting physiotherapy engagement\" would better reflect the proof-of-concept, technical nature of the contribution."
Response:
We agree completely with this suggestion. The original title implied demonstrated clinical improvements which we have not evaluated. We have revised the title to accurately reflect the proof-of-concept nature of the work.
Changes Made:
Location: Title
Change: Revised to reflect technical contribution
ORIGINAL: "ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Enhancing Physiotherapy Adherence in School-Age Children (5-18 Years) with Congenital Talipes Equinovarus" REVISED: "ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Physiotherapy Exercise Monitoring in School-Age Children with Congenital Talipes Equinovarus"

COMMENT 2: Pseudo-Replication Clarification and Per-Participant Performance
Reviewer Comment (Verbatim):
"The confusion matrix values appear to reflect predictions on augmented samples rather than independent test trials. This is pseudo-replication. While data augmentation is reasonable given the small dataset, the manuscript should clarify that these \"instances\" are synthetic variants of a small number of original trials and do not represent hundreds of independent observations. Please also Report per-participant or per-original-trial performance to provide a more interpretable measure of generalization."

Response:
We thank the reviewer for this important clarification. The reviewer is correct that the confusion matrix values represent augmented samples, not independent observations. We have now:
(1) Added explicit clarification that the test set comprises 36 original trials (6 participants × 6 exercises) augmented to 3,600 samples; (2) Added a new Table (Table 2) reporting per-participant classification accuracy; (3) Added per-original-trial accuracy using majority voting; (4) Added a new figure (Figure 14) showing per-participant and per-exercise performance.
Key clarification added: The test set consisted of 36 original trials from 6 held-out participants, augmented to 3,600 samples. Per-participant accuracy ranged from 94.2% to 97.8% (mean: 95.9% ± 1.2%). At the original trial level (majority vote across augmented samples), 35/36 trials (97.2%) were correctly classified.
Changes Made:
Location: Section 2.4.3 (Model Training and Validation)
Change: Added clarification about augmentation and sample counts
Added: "It is important to note that the test set consisted of 36 original exercise trials (6 participants × 6 exercises each) that were augmented 100-fold to produce 3,600 evaluation samples. The values in the confusion matrix therefore represent predictions on augmented samples derived from these 36 original trials, not 3,600 independent observations. This augmentation was applied identically to training and test sets to maintain consistency, but readers should interpret the large sample counts as synthetic variants of a limited number of real trials."
Location: Section 3.1 - New subsection 3.1.7
Change: Added per-participant performance analysis
New subsection "3.1.7 Per-Participant and Per-Trial Performance" with Table 2 showing individual participant accuracies and Figure 14 showing performance breakdown.

COMMENT 3: Vancouver Reference Numbering
Reviewer Comment (Verbatim):
"Please number references in ascending order in the text per the Vancouver formatting system. Then, present a numbered References section where the numbers correspond to those in the text."
Response:
We have converted all in-text citations from author-year format to numbered Vancouver style. References are now numbered in order of first appearance in the text, and the References section presents them in numerical order.
Changes Made:
Location: Throughout manuscript
Change: Converted to Vancouver numbered citations
Example changes: - "(Smythe, 2023)" → "(1)" - "(Pavone et al., as cited in Demographic study, 2024)" → "(2)" - "(Dobbs et al., 2004; Haft et al., 2007)" → "(3,4)" All references have been renumbered in order of first appearance.
Location: References section
Change: Converted to numbered list format
References are now presented as: [1] Smythe T, Mutasa B, Goulios A, Lavy C. The global birth prevalence of clubfoot... [2] Pavone V, et al. ... [3] Dobbs MB, Rudzki JR, Purcell DB, et al. ... etc.

COMMENT 4: APA Reference Formatting
Reviewer Comment (Verbatim):
"Present References in APA format. When the number of authors is > 6, list the first 6 followed by an et al. When the # of authors is 6 or less, please list all the authors."
Response:
We have reformatted all references to APA 7th edition style with the specified author listing rules (all authors if ≤6, first 6 + et al. if >6). Note: We have combined this with the Vancouver numbered system as requested in Comment 3, resulting in numbered references in APA format.

Changes Made:
Location: References section
Change: Reformatted to APA style with author rules
Example (6 or fewer authors - list all): [1] Smythe, T., Mutasa, B., Goulios, A., & Lavy, C. (2023). The global birth prevalence of clubfoot: A systematic review and meta-analysis. EClinicalMedicine, 58, 101980. https://doi.org/10.1016/j.eclinm.2023.101980 Example (more than 6 authors - first 6 + et al.): [15] Zhang, W., Liu, Q., Chen, X., Wang, Y., Smith, J., Johnson, R., et al. (2024). Promoting child and adolescent health through wearable devices. International Journal of Environmental Research and Public Health, 21(3), 268.

COMMENT 5: Remove Opinionated Language
Reviewer Comment (Verbatim):
"Purge all opinionated phrase and replace with what the data support. For example, replace \"..such high performance…\", \"….which is also an excellent score……\", \"…highly accurate….. \"."
Response:
We have systematically reviewed the manuscript and replaced all subjective/opinionated language with objective, data-supported statements. A comprehensive list of changes is provided below.
Changes Made:

Original (Opinionated)	Revised (Objective)
"such high performance"	"an accuracy of 95.92%"
"which is also an excellent score"	"an AUC of 0.995"
"highly accurate"	"achieved 95.92% classification accuracy"
"very strong performance"	"accuracy above 94%"
"extremely reliable"	"achieved AUC values of 1.00 for five classes"
"excellent results"	"correction rates exceeding 90%"
"amazing success rate"	"reported success rates above 90%"
"very promising"	"demonstrated technical feasibility"
"really big problem"	"a significant public health concern"
"very important"	"critical for maintaining correction"

COMMENT 6: Discussion of Platform-Agnostic Design Limitation
Reviewer Comment (Verbatim):
"To me, it seems highly doubtful; that - in an age of smart phones and handheld game consoles - children would voluntarily subject themselves to this rudimentary chunky gaming device without considerable parental pressure. As I see it, an add-on-universal-app for children's smart phones or gaming consoles would be better suited. For example, a child would not be able to play ANY game in a certain time-window unless that particular game were to be 'linked' to completion of these exercises - upon which the device would be unlocked and the child could play games as usual. The 'game layer' of your invention should actually have been modular and platform-agnostic and represents a missed opportunity and a design limitation. Perhaps you will improve in this regard as you iterate your device. Please provide some discussion on this angle of thought."
Response:
We appreciate this thoughtful critique and agree that a platform-agnostic, modular approach would be superior to our current standalone game implementation. The reviewer's suggestion of integrating exercise completion as a gatekeeper for existing preferred games/apps is compelling and would likely improve real-world adoption. We have added a substantive discussion of this limitation and future direction.
Changes Made:
Location: Section 3.2 - New subsection 3.2.5 "Design Limitations and Future Directions"
Change: Added discussion of platform-agnostic design

"3.2.5 Design Limitations and Future Directions for the Gamification Approach A significant limitation of our current implementation is that the Varus game is a standalone, purpose-built application rather than a platform-agnostic system. In an era where children have access to sophisticated smartphones, tablets, and gaming consoles with extensive game libraries, a simple space-shooter game may have limited appeal compared to children's preferred entertainment options. A more compelling approach, which we did not implement but recommend for future development, would be a modular 'exercise gatekeeper' system. In this model, the ToeFro sensor would communicate with a background service on the child's device (smartphone, tablet, or gaming console). This service would lock access to ALL games or entertainment apps during designated physiotherapy time windows. Completing the prescribed exercises, as verified by the sensor and ML model, would unlock the device for normal use. This approach leverages the child's existing motivation to access their preferred games rather than requiring them to engage with a potentially less appealing dedicated application. Such a system would require: (a) development of platform-specific background services for iOS, Android, and gaming platforms; (b) parental control integration; (c) configurable time windows and exercise prescriptions; and (d) appropriate safeguards against circumvention. While technically more complex than our current proof-of-concept, this approach would likely achieve substantially higher voluntary engagement and represents an important direction for future iterations of this work."

COMMENT 7: Code, Data, and Video Availability
Reviewer Comment (Verbatim):
"Please make your code and data available on a public depository such as GitHub. Provide a link to your deposited data in the manuscript. Also provide links as supplementary files to videos that actually show the 6 exercises presented in the paper. YouTube videos are acceptable. The reader should have a general visual idea of how these exercises are performed."
Response:
We have made the code, data, and demonstration video publicly available. The GitHub repository contains the Arduino/ESP32 firmware, Python data processing and ML training scripts, Unity game source code, and anonymized sensor data. The YouTube video demonstrates all six exercises used in the study.
Resources Made Available:
GitHub Repository: https://github.com/sankar-mechengg/ToeFro-Varus
Repository contents: ESP32 firmware, Python ML pipeline, Unity game source, anonymized dataset, hardware schematics
Exercise Demonstration Video: https://www.youtube.com/watch?v=KeX06cie1pA
Video contents: Demonstration of all six physiotherapy exercises (Frog Jump, Kangaroo Jump, Heel Walking, Tip Toe Walking, Sidekick Run, Scissor Run) performed with the ToeFro device
Changes Made in Manuscript:
Location: End of Section 2 (Materials and Methods)
Change: Added Data and Code Availability statement
"2.5 Data and Code Availability All code, anonymized data, and supplementary materials are publicly available. The GitHub repository (https://github.com/sankar-mechengg/ToeFro-Varus) contains: (a) ESP32 firmware for the ToeFro sensor, (b) Python scripts for data processing, feature extraction, and machine learning model training, (c) Unity project files for the Varus game, (d) anonymized sensor data from all participants, and (e) hardware schematics and 3D printing files for the enclosure. A video demonstrating all six physiotherapy exercises is available at https://www.youtube.com/watch?v=KeX06cie1pA."

SUMMARY OF REVISIONS (ROUND 3)

Comment	Action Taken	Location
1	Title softened to "monitoring"	Title
2	Clarified pseudo-replication; added per-participant	Sections 2.4.3, 3.1.7, Table 2, Fig

	table	14
3	Converted to Vancouver numbered citations	Throughout; References
4	Reformatted references to APA style	References section
5	Replaced opinionated language with objective statements	Throughout (10+ instances)
6	Added discussion of platform-agnostic design limitation	New Section 3.2.5
7	Added GitHub and YouTube links	New Section 2.5

We believe these revisions fully address the reviewer's comments. The manuscript now has a more accurate title, properly contextualized results, correctly formatted references, objective language, transparent discussion of design limitations, and publicly available code and data. We thank the reviewer for their constructive feedback throughout this process, which has substantially improved the manuscript.

Sincerely,

Varun Nadkarni and Sankar Balasubramanian

Thank you for addressing my comments. The manuscript is now much improved. Some additional minor inconsistencies remain.

1.Section 3.2.1 asserts that the system can count repetitions and duration, which contradicts Section 2.2.2 stating partial exercise detection/repetition counting was not implemented—this overstatement should be corrected.

2.Even with softened language, the paper still implicitly suggests rehabilitation utility without measuring any clinical outcome. Please remove all overreach such as , "supports physiotherapy engagement", "can assist in rehabilitation workflows", "may improve home-based physiotherapy monitoring", "enables more effective physiotherapy supervision", "has potential to improve physiotherapy outcomes". Replace with suggested phrases such as ""provides automated monitoring of prescribed physiotherapy exercises", "enables objective tracking of physiotherapy exercise execution", "facilitates real-time recognition of prescribed exercise types", "can be integrated into digital physiotherapy monitoring systems", "can support data collection within rehabilitation research or pilot deployments", "is technically compatible with digital rehabilitation monitoring pipelines", "enables remote supervision through automated exercise classification", "has potential to be evaluated in future studies for associations with physiotherapy outcomes"…. etc. Please check the entire manuscript.

3.You report high accuracy with 6 participants. This is too less for generalization. Please add "While performance is high within this small cohort, the dataset is too limited to support claims of generalization across the pediatric CTEV population."

4.You have not compared to baselines or simpler models. Therefore, it is not clear why ML + gamification is necessary versus many simpler threshold-based heuristics or clinician labeling. Include something like "The present study does not claim that ML-based classification or gamification is superior to simpler threshold-based or clinician-labeled approaches; rather, it demonstrates the technical feasibility of an automated, extensible framework that could be compared against such baselines in future work."

RESPONSE TO REVIEWER COMMENTS

Revision Round 5

Manuscript: ToeFro-Varus: A Gamified, Machine Learning-Powered Wearable for Physiotherapy Exercise Monitoring in School-Age Children with Congenital Talipes Equinovarus

Journal of High School Science | Date: 18 February 2026

We thank the reviewer for their continued careful attention to precision in our claims. We have addressed all four comments by: (1) correcting the internal contradiction regarding repetition counting, (2) systematically replacing all language implying clinical utility with technically accurate

phrasing, (3) adding an explicit statement about generalization limitations, and (4) adding a statement clarifying that we do not claim superiority over simpler approaches. All changes are detailed below.

COMMENT 1: Contradiction Regarding Repetition Counting

Reviewer Comment (Verbatim):

"Section 3.2.1 asserts that the system can count repetitions and duration, which contradicts Section 2.2.2 stating partial exercise detection/repetition counting was not implemented—this overstatement should be corrected."

Response:

We thank the reviewer for identifying this internal contradiction. The reviewer is correct: Section 2.2.2 explicitly states that repetition counting was not implemented, and therefore Section 3.2.1 should not claim this capability. We have corrected Section 3.2.1 to remove the erroneous claim and ensure consistency with the Methods section.

Changes Made:

Location: Section 3.2.1

Change: Removed false claim about repetition counting

ORIGINAL: "The system can count repetitions and track exercise duration, providing physiotherapists with detailed session data." REVISED: "The system classifies exercise types in real-time, which could be extended in future implementations to include repetition counting and duration tracking. The current prototype validates exercise completion but does not count individual repetitions."

We also performed a full manuscript search for any other mentions of "repetition," "counting," or "duration" to ensure no other contradictions exist. No additional inconsistencies were found.

COMMENT 2: Remove Language Implying Clinical Utility

Reviewer Comment (Verbatim):

"Even with softened language, the paper still implicitly suggests rehabilitation utility without measuring any clinical outcome. Please remove all overreach such as, \"supports physiotherapy engagement\", \"can assist in rehabilitation workflows\", \"may improve home-based physiotherapy monitoring\", \"enables more effective physiotherapy supervision\", \"has potential to improve physiotherapy outcomes\". Replace with suggested phrases such as \"provides automated monitoring of prescribed physiotherapy exercises\", \"enables objective tracking of physiotherapy exercise execution\", \"facilitates real-time recognition of prescribed exercise types\", \"can be integrated into digital physiotherapy monitoring systems\", \"can support data collection within rehabilitation research or pilot deployments\", \"is technically compatible with digital rehabilitation monitoring pipelines\", \"enables remote supervision through automated exercise classification\", \"has potential to be evaluated in future studies for associations with physiotherapy outcomes\".... etc. Please check the entire manuscript."

Response:

We accept this criticism entirely. We have systematically searched the entire manuscript for any language that implies clinical utility or therapeutic benefit and replaced it with technically accurate phrasing that describes only what the system demonstrably does. The complete list of replacements is provided below.

Comprehensive List of Replacements:

REMOVED (Implies Clinical Utility)	REPLACED WITH (Technical Description)
supports physiotherapy engagement	provides automated monitoring of prescribed physiotherapy exercises
can assist in rehabilitation workflows	can be integrated into digital physiotherapy monitoring systems
may improve home-based physiotherapy monitoring	enables remote supervision through automated exercise classification

enables more effective physiotherapy supervision	enables objective tracking of physiotherapy exercise execution
has potential to improve physiotherapy outcomes	has potential to be evaluated in future studies for associations with physiotherapy outcomes
enhance patient motivation	provides gamified feedback upon exercise completion
improve long-term outcomes	facilitate data collection for future outcome studies
help ensure better rehabilitation	facilitate real-time recognition of prescribed exercise types
make physiotherapy more effective	is technically compatible with digital rehabilitation monitoring pipelines
support the rehabilitation process	can support data collection within rehabilitation research or pilot deployments
contribute to better treatment adherence	provide objective exercise completion data
helps children stay motivated	provides game-based feedback contingent on exercise classification

Sections Modified:
• Abstract (2 instances) • Section 1.4 Novel Approach (3 instances) • Section 1.5 Study Objectives (1 instance) • Section 3.2.1 Efficacy for Remote Monitoring (4 instances) • Section 3.2.2 Gamification Benefits (2 instances) • Section 4 Conclusion (3 instances)

COMMENT 3: Generalization Limitation Statement
Reviewer Comment (Verbatim):
"You report high accuracy with 6 participants. This is too less for generalization. Please add \"While performance is high within this small cohort, the dataset is too limited to support claims of generalization across the pediatric CTEV population.\""
Response:
We agree completely. We have added the reviewer's suggested statement verbatim, and have also added similar caveats in the Abstract and Results sections to ensure this limitation is clearly communicated throughout the manuscript.
Changes Made:
Location: Section 3.2.4 (Limitations) – First paragraph
Change: Added verbatim statement
Added: "While performance is high within this small cohort, the dataset is too limited to support claims of generalization across the pediatric CTEV population."
Location: Abstract – After reporting accuracy
Change: Added generalization caveat
Added: "However, these results are based on a small test cohort (n=6 participants) and should not be interpreted as evidence of generalization to the broader pediatric CTEV population."
Location: Section 3.1.1 – After reporting Gradient Boosting results
Change: Added inline caveat
Added: "It must be noted that this performance was measured on only 6 test participants, which is insufficient to claim generalization."

COMMENT 4: No Comparison to Simpler Baselines
Reviewer Comment (Verbatim):
"You have not compared to baselines or simpler models. Therefore, it is not clear why ML + gamification is necessary versus many simpler threshold-based heuristics or clinician labeling. Include something like \"The present study does not claim that ML-based classification or gamification is superior to simpler threshold-based or clinician-labeled approaches; rather, it demonstrates the technical feasibility of an automated, extensible framework that could be compared against such baselines in future work.\""
Response:

We agree that we have not demonstrated superiority over simpler approaches, and we should not imply this. We have added the reviewer's suggested statement and have also added a related statement in the Discussion section.

Changes Made:

Location: Section 3.2.4 (Limitations)

Change: Added verbatim statement

Added: "The present study does not claim that ML-based classification or gamification is superior to simpler threshold-based or clinician-labeled approaches; rather, it demonstrates the technical feasibility of an automated, extensible framework that could be compared against such baselines in future work."

Location: Section 4 (Conclusion) – Second-to-last paragraph

Change: Added clarification

Added: "We emphasize that this work demonstrates technical feasibility of an automated approach, not superiority over alternative methods such as threshold-based heuristics, rule-based systems, or direct clinician observation. Comparative studies evaluating the relative merits of different monitoring approaches remain an important direction for future research."

SUMMARY OF REVISIONS (ROUND 4)

Comment	Action Taken	Location
1	Corrected contradiction re: repetition counting	Section 3.2.1
2	Replaced 12+ instances of clinical utility language	Throughout (see table)
3	Added generalization limitation statement (verbatim)	Abstract, 3.1.1, 3.2.4
4	Added baseline comparison caveat (verbatim)	Sections 3.2.4, 4

We believe these revisions fully address the reviewer’s comments. The manuscript now contains no claims of clinical utility, acknowledges the generalization limitations explicitly, and clarifies that we do not claim superiority over simpler approaches. We thank the reviewer for their rigorous attention to precision, which has significantly improved the accuracy of our claims.

Sincerely,

Varun Nadkarni and Sankar Balasubramanian

Thank you for addressing my comments. Accepted. However, BEFORE we begin copyediting, please rescale figure 12 so that the Y axis is between 80 and 100, rescale figure 14 (both panels) so the they Y axis is between 90 and 100. Please also label figure 5 components. You can attach these to a discussion post and send them.

1.send Figure 12 as a separate JPEG file with the Y axis between 80 and 100.

2.Figure 14 as a separate JPEG file with the Y axis between 90 and 100 (both left and right panels)

3.Figure 5 as a separate JPEG file with all components clearly labeled in the figure.
