

Article

From Data to Autonomy: Integrating Demographic Factors and AI Models for Expert-Free Exercise Coaching

Uğur Özbalkan ^{1,2}  and Özgür Can Turna ^{1,*} ¹ Department of Computer Engineering, İstanbul University-Cerrahpaşa, İstanbul 34320, Türkiye; ozbalkanugur95@gmail.com or ugur.ozbalkan@ibu.edu.tr² Department of Computer Engineering, Fenerbahçe University, İstanbul 43020, Türkiye

* Correspondence: ozgurcan.turna@iuc.edu.tr

Abstract

This study investigates the performance of three deep learning architectures—LSTM with Attention, GRU with Attention, and Transformer—in the context of real-time, self-guided exercise classification, using coordinate data collected from 103 participants via a dual-camera system. Each model was evaluated over ten randomized runs to ensure robustness and statistical validity. The GRU + Attention and LSTM + Attention models demonstrated consistently high test accuracy (mean $\approx 98.9\%$), while the Transformer model yielded significantly lower accuracy (mean $\approx 96.6\%$) with greater variance. Paired *t*-tests confirmed that the difference between LSTM and GRU models was not statistically significant ($p = 0.9249$), while both models significantly outperformed the Transformer architecture ($p < 0.01$). In addition, participant-specific features, such as athletic experience and BMI, were found to affect classification accuracy. These findings support the feasibility of AI-based feedback systems in enhancing unsupervised training, offering a scalable solution to bridge the gap between expert supervision and autonomous physical practice.

Keywords: artificial intelligence in exercise monitoring; deep learning for exercise feedback; human motion tracking; real-time movement classification; self-guided fitness training

1. Introduction

Artificial intelligence (AI) is undergoing a paradigm shift in sports applications, where techniques like performance analysis, training optimization, and decision support systems are increasingly being utilized. AI techniques such as machine learning, computer vision, and pattern recognition are utilized to analyze athlete movements, classify sports activities, and provide personalized training strategies [1,2]. The utilization of image processing and video analysis tools, such as Coach's Eye and Dartfish, has been validated to enhance skill assessment and feedback [3]. Recent innovations in sports analytics include AI-driven systems that enhance refereeing accuracy and enable precise injury prediction [4]. In weight training, AI applications have shown promising results in evaluating exercise techniques [5]. However, challenges persist in data collection, AI controllability, and result interpretability [1]. Despite these challenges, AI holds significant potential for biomechanical analysis and performance optimization [6,7].

AI-based feedback systems have shown promise in improving user motivation and adherence compared to traditional methods. Studies indicate that AI feedback enhances movement accuracy, balance, and health outcomes in older adults practicing tai chi [8]. Computerized feedback systems have been found to increase adherence to exercise regimes



Academic Editor: Ephraim Suhir

Received: 31 October 2025

Revised: 21 December 2025

Accepted: 30 December 2025

Published: 3 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and reduce dropout rates [9]. Real-time AI-assisted posture correction has been shown to improve exercise technique [10], while AI-generated personalized fitness regimes have the potential to improve physical and mental well-being [11]. Pose prediction-based feedback has been demonstrated to enhance motivation and performance in home-based exercises [12]. AI-driven personalization of social comparison goals has demonstrated small to moderate effects on increasing physical activity motivation [13]. Furthermore, the integration of behavior change techniques into AI fitness applications has been identified as crucial to encourage user engagement and facilitate behavior change [14].

This study examines the effectiveness of three attention-enhanced deep learning architectures—LSTM, GRU, and Transformer—for improving the accuracy of self-guided exercise execution. Grounded in the premise that recurrent networks encode temporal continuity and phase transitions more explicitly than purely self-attentional models, we hypothesize that the LSTM + Attention configuration will yield superior real-time motion classification by leveraging structured memory dynamics. Our framework operationalizes this hypothesis within a pose-based temporal pipeline and benchmarks model behavior not only on average accuracy but also on stability across runs and confusion patterns among visually similar movements, thereby providing a rigorous basis for architectural comparison.

Beyond predictive performance, the work targets a central challenge in autonomous physical training: mitigating injury risk stemming from improper technique. To that end, the system is designed to deliver automated, immediate feedback aligned with salient phases of each repetition, enabling users to adjust form *in situ* rather than retrospectively. By linking model outputs to interpretable cues (e.g., range of motion, tempo adherence, joint alignment), the approach aims to support safer and more effective practice in both athletic and everyday health contexts, while laying groundwork for personalized coaching via contextual covariates (e.g., experience level, anthropometrics) and adaptable decision thresholds.

2. Literature Review

Recent research highlights the potential of computer vision and AI in delivering real-time feedback for physical exercises, reducing reliance on human coaches. These systems leverage pose estimation algorithms like YOLOv7-pose and PoseNet to track body keypoints and analyze posture [15,16]. Machine learning techniques are employed to compare user performance with expert demonstrations, enabling immediate posture correction and repetition counting [17,18]. Some systems offer 3D reconstruction of human motion and personalized feedback based on predefined standards [19]. These approaches show notable improvements in the safety and efficacy of activities like weightlifting, yoga, and fitness routines [20,21]. User studies have demonstrated positive responses to these AI-powered systems, suggesting their potential to enhance workout effectiveness and motivation, particularly in circumstances where professional guidance is unavailable [10].

Deep learning models have demonstrated considerable potential in the domains of exercise performance assessment and injury prevention in sports. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs) demonstrate strengths in image recognition and time-series analysis for injury prediction [22]. A hybrid CNN-BiGRU-CBAM model has achieved state-of-the-art results in recognizing sports and daily activities using wearable sensor data [23]. Machine learning techniques such as random forests and artificial neural networks have improved injury risk assessments [24]. Deep learning approaches have been applied to rehabilitation, including a redefined prairie dog optimized Bi-LSTM for personalized recovery plans [25] and a Hybridized Hierarchical Deep CNN for exercise rehabilitation [26]. Support Vector Machines have been widely used in both IMU and vision-based studies [27]. Researchers

have developed deep learning frameworks to assess rehabilitation exercises [28] and monitor on-field player workload exposure [29].

The potential of artificial intelligence (AI) and image processing technologies to enhance athletes' performance through automated movement correction and feedback has been well demonstrated in numerous studies [5,30]. These technologies analyze sports postures with high accuracy, deliver real-time feedback, and assess exercise techniques. AI-driven systems can evaluate human body poses, support biomechanical studies, and identify technical flaws that may increase injury risk [31]. Machine learning and deep learning techniques enable efficient data analysis, decision support, and the creation of personalized training plans [32]. Furthermore, the integration of motion capture systems with artificial intelligence enables automated qualitative assessments by extracting movement errors from kinematic data [33]. In addition, the combination of wearable technologies with artificial intelligence has been shown to process physiological and kinematic data, for performance optimization and injury prediction [34]. Furthermore, neural network-based systems have been shown to provide real-time visual feedback for the correction of inaccurate postures and motions [35].

3. Materials and Methods

This section details the data collection process, AI model architectures, and experimental protocols designed to evaluate self-guided exercise accuracy.

3.1. Data Collection Software and Procedure

The present study focuses on the recognition of movements by artificial intelligence. In this context, software has been developed with Python 3.9 for the purpose of data collection, with the objective of creating a dataset. The software's operational sequence is delineated as follows:

- ❖ Software Initialization: Python libraries (e.g., MediaPipe, OpenCV) and variables were initialized.
- ❖ Participant Registration: Demographic data (weight, height, gender, etc.) were collected via a GUI.

All recordings were conducted in the Sports Sciences Laboratory of Fenerbahçe University under controlled environmental conditions. At the beginning of the custom-developed software workflow, participants entered their demographic information (age, weight, height, BMI, gender, sports experience) directly into the system via a standardized GUI form. This ensured that all demographic data were collected consistently before recording. The laboratory setup was arranged to guarantee that each participant remained fully visible from head to toe during all movements.

- ❖ Camera Setup: Dual cameras (computer + tablet) were calibrated at 30 FPS.
- ❖ Exercise Selection: Participants selected exercises via keyboard inputs (e.g., "a" for squats).
- ❖ Data Acquisition: Real-time skeleton coordinates (33 landmarks) were extracted using MediaPipe [Version 0.10.10] Pose. Video recording commenced concurrently with data collection.
- ❖ Data Processing: Skeleton poses were identified, normalized, and processed to ensure data consistency.
- ❖ Exercise Completion: Participants completed the exercises, triggering the end of data acquisition. Libraries (e.g., MediaPipe, OpenCV) and variables were initialized.
- ❖ Data Storage: Normalized x/y coordinates and scores were systematically saved to Excel files.

- ❖ **Software Termination:** The program was properly terminated, ensuring all resources were safely released and files closed.

In the dataset, the information of the following headings is received and calculated:

- ID (It increases automatically as the information of the movement number.)
- Movement Name (Exercise Type)
- Time (incremental time information in 0.033 s intervals)
- Weight
- Height
- Gender
- Age (Calculated with Date of Birth information)
- Sports Experience (in years)
- Sporting Level (Beginner/Intermediate/Advanced)
- Body Mass Index calculated from weight and height information
- Distance (calculated as (Actual face width $\times$ Focal length)/Face width in the image)

The estimation of distance was conducted on a single occasion at the commencement of each recording. In the present study, real-time depth compensation was not a component of the research design. It was established that, due to the standardization of all 2D pose coordinates to a [0, 1] range, minor forward–backward movements exerted negligible influence on the temporal patterns employed by the classification models. Nevertheless, the absence of real-time depth monitoring is acknowledged as a limitation and will be addressed in future work through the use of depth sensors or continuous scale tracking.

- x and y coordinate information of 33 points of the body taken from each of the 2 cameras (calculated by MediaPipe)

The selection of 33 anatomical landmarks is informed by the pre-defined keypoint configuration of the MediaPipe Pose model. The selection of this model was made on the basis of its provision of a lightweight, real-time inference pipeline with high temporal stability across frames, a property which is essential for synchronized dual-camera acquisition. In comparison with other pose estimators, such as OpenPose and HRNet, MediaPipe Pose exhibited superior robustness and reduced latency during the pilot tests. This renders it well-suited for laboratory data collection and temporal sequence modeling.

Unlike prior single-camera pose-based exercise datasets, this study introduces a synchronized dual-camera temporal fusion pipeline designed to reduce viewpoint dependency and improve temporal smoothness of pose trajectories. Each frame pair is time-aligned using software-based latency compensation, and the combined sequences provide more stable movement dynamics for downstream temporal modeling. This dual-view synchronized acquisition constitutes a methodological contribution beyond standard MediaPipe-based pipelines and serves as the foundation for the model comparison framework presented in later sections.

3.2. Coordinate Validation and Data Collection Process for Multi-Camera Pose Estimation

The dual-camera system consisted of a 30 FPS tablet camera and a 30 FPS computer webcam, both operating under fixed laboratory lighting. Recordings were carried out using a computer equipped with a 13th Gen Intel® Core™ i9-13900H processor (2.60 GHz) (Intel, Santa Clara, CA, USA) and 16 GB RAM. Cameras were positioned to ensure that the participant's entire body—including the feet—remained fully visible in the frame. Since MediaPipe Pose extracts 33 anatomical landmarks directly from standard RGB frames, participants were not required to wear specific clothing; no occlusion or contrast issues were observed during data collection. With the prepared software, 2 cameras are activated by taking the desired information from the person who will perform the movement and

calibrating with the camera. With each of the 2 cameras, 33 points of the body are shown on the screen. The movement starts by pressing the following keys on the keyboard and x and y coordinate information of 33 points are obtained from both cameras at 0.033 s intervals depending on time.

- a Free-Weight Squat
- b Dumbbell Biceps Curl
- c Dumbbell Lateral Raise
- d Standing Calf Raise
- e Dumbbell Shoulder Press
- f Terminating the movement

MediaPipe is a versatile framework for developing perception applications, addressing challenges in processing perceptual inputs across various devices [36]. MediaPipe Pose, a versatile machine learning framework for pose estimation, was utilized to identify 33 specific anatomical landmarks on the human skeleton from video frames. These landmarks include anatomical points such as shoulders, elbows, wrists, hips, knees, and ankles, and are detected through a neural network. Specifically, the MediaPipe Pose model was configured in Python using the `mp.solutions.pose` library with parameters: `static_image_mode = False`, `model_complexity = 2`, and confidence thresholds (`min_detection_confidence = 0.5`, `min_tracking_confidence = 0.5`).

A dual-camera setup (computer and tablet) captured synchronized videos simultaneously at 30 frames per second (FPS). To monitor and correct inter-camera latency, timestamps were recorded using the `time.time()` function in Python, and frame delays between the two cameras were computed. Although hardware-based synchronization (e.g., GPIO triggers) was not implemented in this prototype, software-based latency tracking was used to minimize temporal misalignment in post-processing.

Data were collected at intervals of 0.033 s, providing high temporal resolution. To maintain synchronization, the latency between the two cameras was calculated in real time using the formula:

$$\text{Latency} = T_{\text{Camera2}} - T_{\text{Camera1}} \quad (1)$$

where each timestamp was recorded at the moment a frame was captured.

For accurate distance estimation between the participant and the camera, focal length was calculated via a one-time calibration process using the formula:

$$\text{Focal Length}(f) = \frac{\text{Width in pixels} \times \text{Known distance}}{\text{Real width of the object}} \quad (2)$$

In this study, an average adult face width of 16.0 cm was assumed as the “real width.” This approximation provided sufficient accuracy for relative distance tracking, but personal face width measurements were not conducted. For higher accuracy in future studies, we plan to compare this approach against ground-truth depth measurements using LiDAR or depth sensors.

The raw landmark coordinates were normalized to a [0, 1] scale to standardize data across varying image resolutions:

$$x_{\text{norm}} = \frac{x_{\text{pixel}}}{\text{image_width}}, y_{\text{norm}} = \frac{y_{\text{pixel}}}{\text{image_width}} \quad (3)$$

- (0, 0) represents the upper left corner,
- (1, 1) represents the bottom right corner.

Prior to the primary data collection phase, a preliminary coordinate validation test was conducted with three subjects to ensure the temporal stability and anatomical coherence of the 33 extracted landmarks. In this validation step, each participant performed five exercises, repeated twice. The resulting time-dependent coordinates were processed to generate stickman representations for visual inspection. This procedure confirmed that the dual-camera system produced synchronized and reliable pose trajectories suitable for subsequent large-scale data acquisition. Figure 1 illustrates the output of this coordinate validation process.

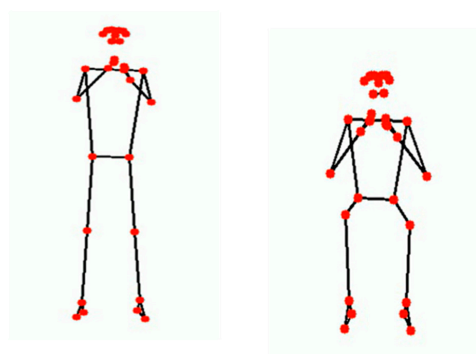


Figure 1. Coordinate Validation Test.

After the testing of the coordinate values the data collection process started. The data collection process was carried out in a laboratory environment with a computer camera and a tablet camera taking 30 frames per second. The data collection area is shown in Figure 2, and the data collection software screen is shown in Figure 3.



Figure 2. Data Collection Area.

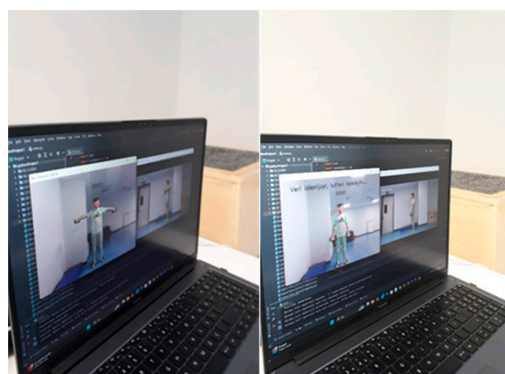


Figure 3. Data Collection Software Screen.

3.3. Model Architecture

The collected dataset was then subjected to the following time-dependent artificial intelligence algorithms, and the accuracy of predicting the movements was tested. In addition to the pose-based temporal sequences, the proposed framework integrates demographic and anthropometric features (e.g., BMI, age, gender, and training experience) through an auxiliary input branch. This design enables the model to simultaneously learn movement dynamics and user-specific physical characteristics, yielding a personalized representation space. Such demographic–pose fusion is largely underexplored in existing exercise-recognition research and constitutes a methodological contribution that extends the architecture beyond conventional pose-only classification pipelines.

- LSTM with Attention Mechanism
- Transformer-Based Motion Recognition
- Gated Recurrent Unit (GRU)

These three architectures were selected because they represent complementary temporal modeling paradigms: LSTM models capture long-term dependencies through gated memory, GRU offers a computationally efficient recurrent alternative, and the Transformer provides a non-recurrent, self-attention-based baseline. Together, these models cover the dominant approaches used in pose-based temporal sequence analysis and enable a balanced comparative evaluation.

3.3.1. LSTM with Attention Mechanism

Recent studies have investigated integrating attention mechanisms with Long Short-Term Memory (LSTM) networks to enhance performance in various sequence-learning tasks. Attention mechanisms improve models' ability to handle long-term dependencies by enabling them to focus on relevant information [37,38].

The stages in creating the model are given in Figure 4 below.

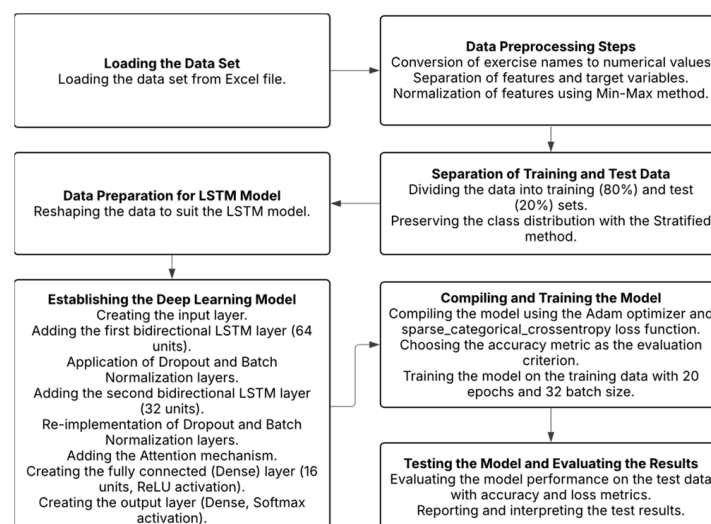


Figure 4. LSTM + Attention Mechanism Model.

3.3.2. Transformer-Based Motion Recognition

Transformers have revolutionized various areas of artificial intelligence, including natural language processing, computer vision and speech processing [39,40]. These models, which utilize self-attention mechanisms, have demonstrated superior performance compared to traditional recurrent neural networks in multiple tasks [41].

The stages in creating the model are given in Figure 5 below.

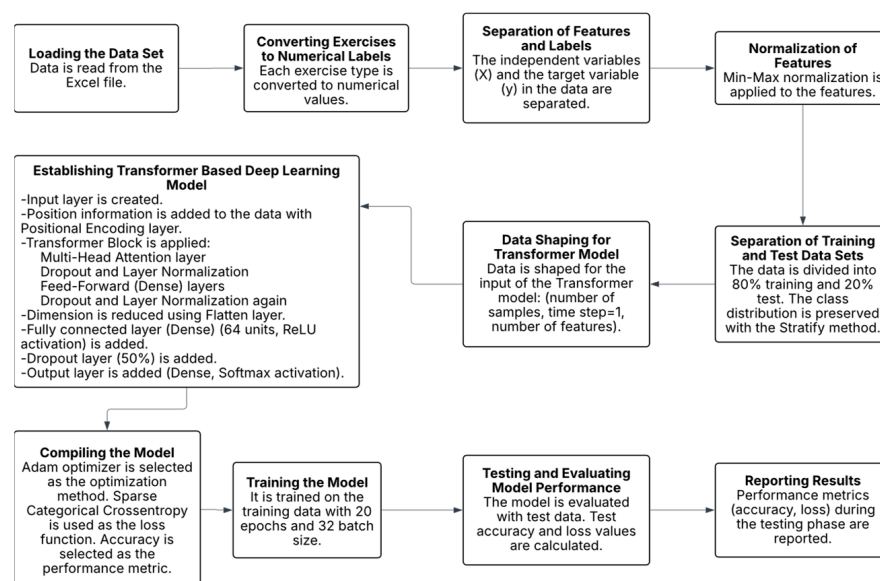


Figure 5. Transformer Model.

3.3.3. GRU with Attention Mechanism

The Gated Recurrent Unit (GRU) algorithm has demonstrated favorable outcomes in a range of applications. For instance, a study by Mohsen [42] reported an accuracy of 97.08% for GRU in human activity recognition. Optimization techniques such as adaptive genetic algorithms [43] and whale optimization algorithms [44] have been employed to enhance the performance of GRU. These studies, along with others of a similar nature, underscore the versatility of GRU and its potential for enhancement through the application of various optimization techniques across diverse domains.

The stages in creating the model are given in Figure 6 below:

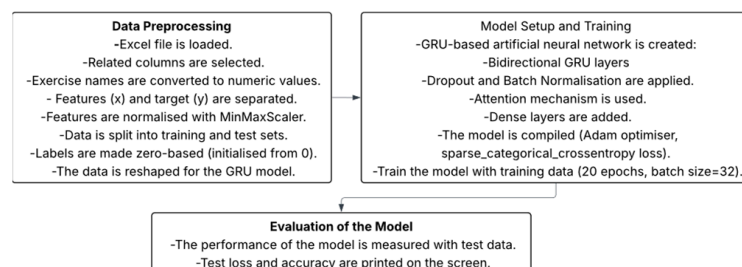


Figure 6. GRU Model.

3.4. Dataset Composition

A total of 103 participants were asked to perform 5 physical movements with 2 repetitions, and time-dependent coordinated information was collected. The statistical information contained in the data set is provided in Tables 1 and 2.

Table 1. Number of Participants by Gender in the Dataset.

Gender	Total Number of Participants
Male Participant	78
Female Participant	25

Table 2. Number of Participants by Level in the Dataset.

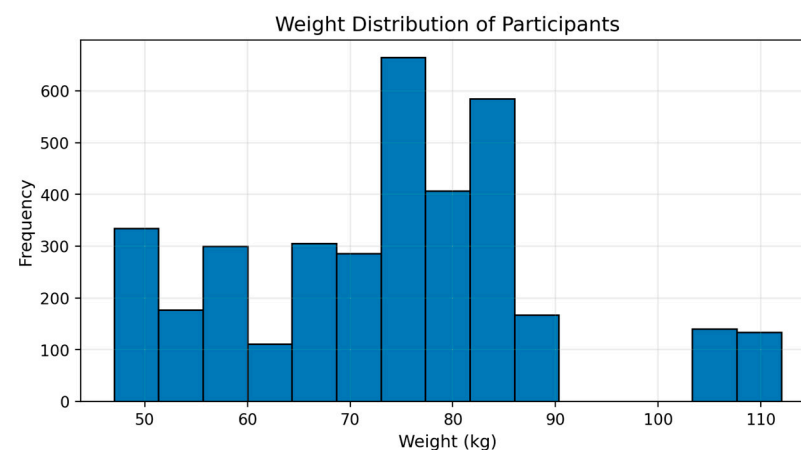
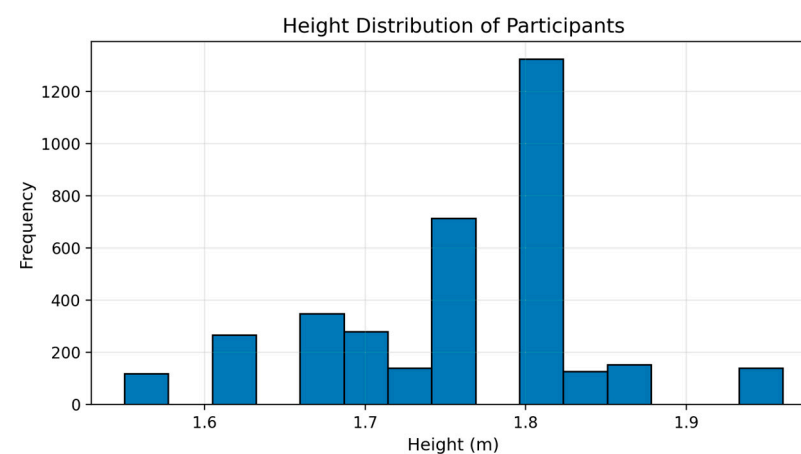
Sporting Level	Total Number of Participants
Beginner	30
Intermediate	47
Advanced	26

- Average Years of Sport Experience: 6.7 years
- Average Weight (kg): 75.13 kg
- Average Height (m): 1.77 m

Mean height and weight values are reported to summarize overall participant characteristics. Accuracy-related analyses were performed using each individual's BMI and experience level rather than aggregated mean values.

Participant Demographic Distributions

The distributions of weight, height, and body mass index (BMI) across the 103 participants are presented in Figures 7–9. These histograms provide a clear overview of the anthropometric diversity within the dataset and illustrate that the participant group spans a wide range of physical characteristics. Such variability enhances the generalizability and applicability of the proposed exercise-recognition system.

**Figure 7.** Histogram of Participants' Weight (kg).**Figure 8.** Histogram of Participants' Height (m).

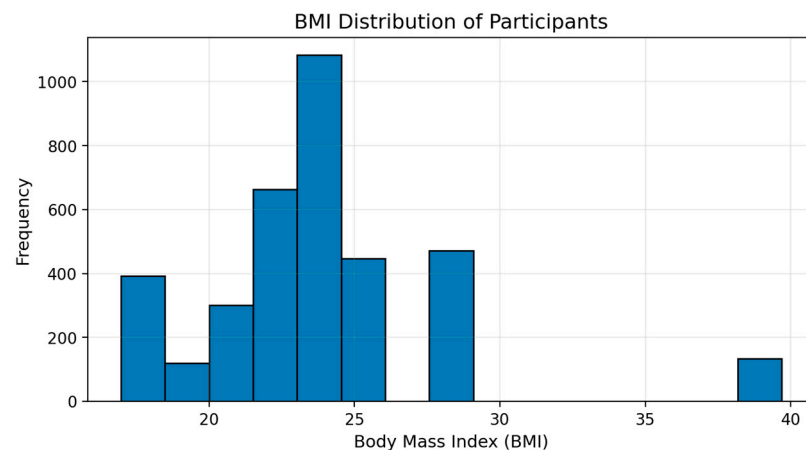


Figure 9. Histogram of Body Mass Index (BMI).

Demographic and anthropometric differences are also relevant for understanding model behavior, as factors such as body proportions and BMI may influence pose trajectories and joint-angle dynamics. Including these distribution plots offers a more comprehensive description of the dataset and supports a deeper interpretation of the model performance reported in later sections.

4. Results

A total of 103 participants were instructed to perform five distinct exercises, each repeated twice, resulting in a comprehensive dataset comprising 2060 samples. The coordinate data of 33 anatomical landmarks were synchronously captured from a dual-camera setup at a constant frame rate using custom-developed software. This data was further enriched with participant-specific information, including demographic attributes and physical characteristics.

To further assess model behavior beyond mean accuracy values, confusion matrices and Receiver Operating Characteristic (ROC) curves were generated for each model. These visualizations provide additional insight into classification consistency and class-specific discrimination ability. Confusion matrices and ROC curves are shown in Figures 10–15.

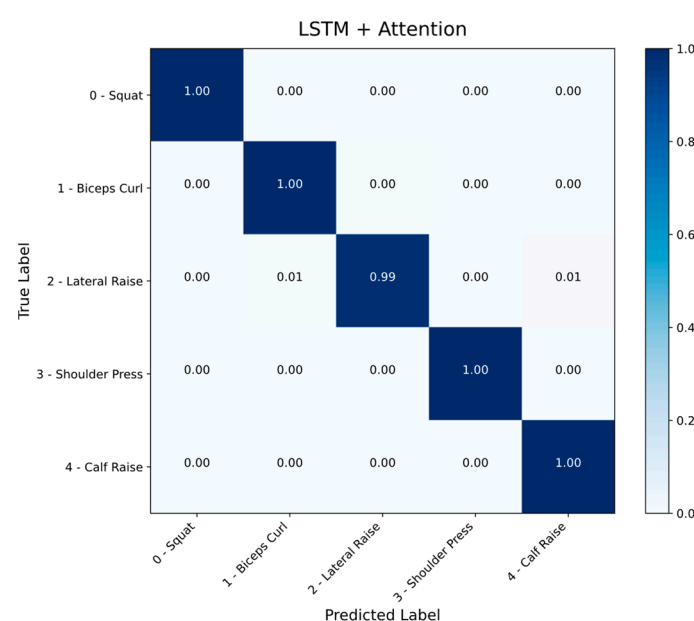


Figure 10. Confusion Matrix—LSTM + Attention.

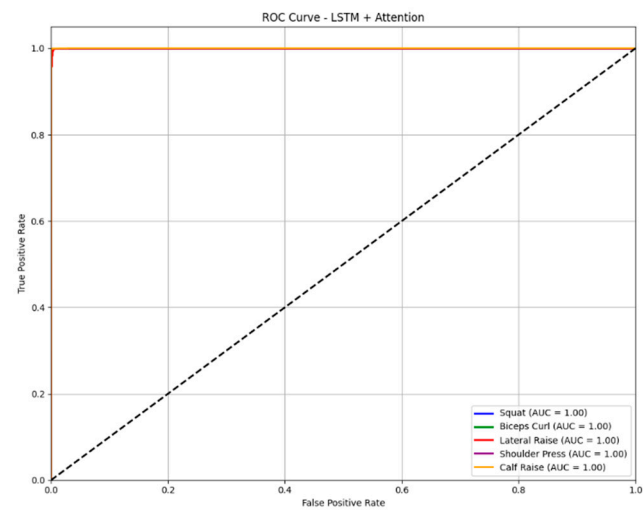


Figure 11. ROC Curve—LSTM + Attention.

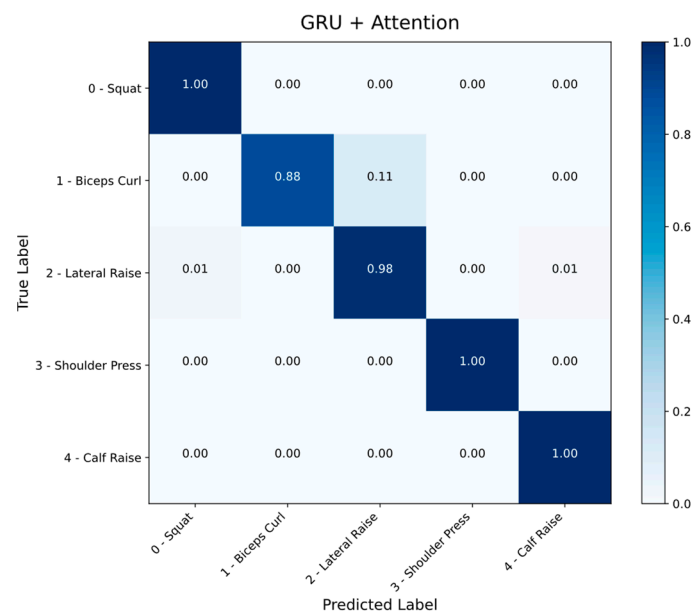


Figure 12. Confusion Matrix—GRU + Attention.

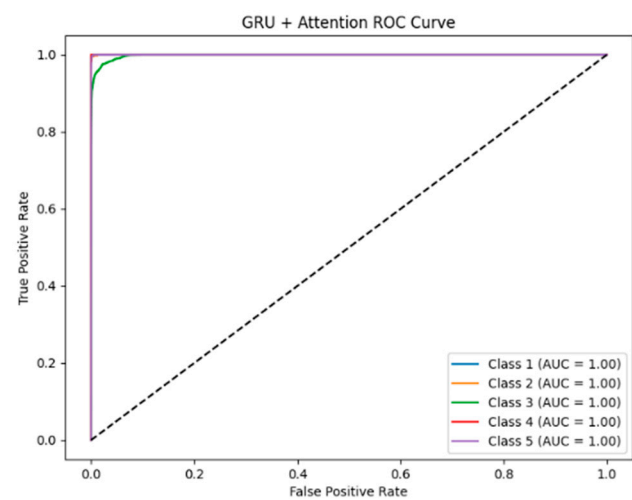


Figure 13. ROC Curve—GRU + Attention.

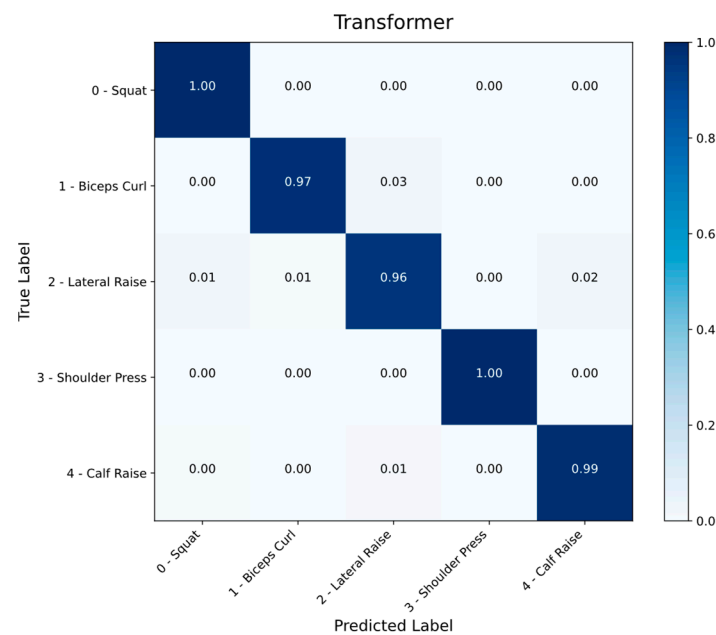


Figure 14. Confusion Matrix—Transformer.

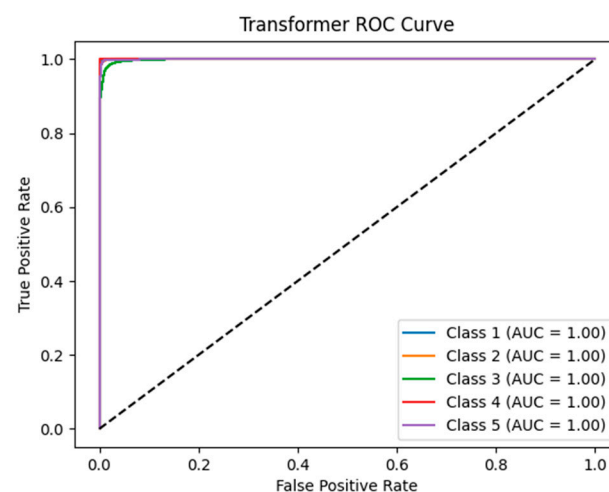


Figure 15. ROC Curve—Transformer.

As shown in the confusion matrices, all three models demonstrated strong overall classification performance, with a clear diagonal dominance indicating high true positive rates. The LSTM + Attention and GRU + Attention models yielded particularly compact and well-defined confusion matrices, reflecting minimal class confusion. In contrast, the Transformer model exhibited relatively more dispersed misclassifications, especially between Biceps Curl and Lateral Raise exercises.

In terms of ROC analysis, all models achieved near-perfect AUC values (close to 1.00) across all classes, indicating excellent sensitivity and specificity. Nevertheless, the ROC curves of the Transformer model revealed slightly more fluctuation, suggesting minor instability in class-level discrimination compared to the recurrent-based models.

- The LSTM + Attention model showed extremely high precision in recognizing “Shoulder Press” and “Calf Raise”, with negligible false positives.
- The GRU + Attention model slightly struggled with distinguishing “Biceps Curl” from “Lateral Raise” yet maintained overall high classification accuracy.

- The Transformer model, despite its high performance, exhibited increased class confusion, as seen in its confusion matrix and the broader, less concave curves of its ROC plot.

The resulting multimodal dataset was then used to train and evaluate various time-dependent artificial intelligence models. Each of the 10 runs corresponds to an independent model training initialized with a different random seed. This approach follows standard practice for robustness and variance assessment, enabling statistically reliable comparison across architectures. Unlike previous works that report single-run results, this study implements a 10-run randomized evaluation protocol for each architecture, followed by paired *t*-tests for statistical significance. This framework contributes a rigorous and reproducible methodology for comparing temporal exercise-recognition models, offering a more reliable assessment than traditional single-seed evaluations. The experimental results obtained from these models are presented as follows:

As shown in Table 3, both the LSTM + Attention and GRU + Attention models produced consistently high and closely aligned accuracy scores across 10 independent runs, with mean accuracy rates of 98.90% and 98.97%, respectively. In contrast, the Transformer-based model achieved a lower average accuracy of 96.57%, and exhibited a higher standard deviation, indicating greater variability and less stable performance.

Table 3. Test Accuracies Across 10 Runs (Per Model).

Run No	LSTM + Attention	GRU + Attention	Transformer
1	0.987	0.9791	0.9488
2	0.9946	0.9964	0.9773
3	0.9929	0.9863	0.982
4	0.996	0.9702	0.9306
5	0.9842	0.9949	0.9842
6	0.9864	0.9979	0.9497
7	0.9848	0.995	0.9895
8	0.9943	0.9931	0.9666
9	0.9939	0.9776	0.9852
10	0.9762	0.9956	0.9387
Mean	0.989	0.9897	0.9657
Standard Deviation	0.0063	0.0092	0.0205

According to the results of the paired *t*-tests presented in Table 4:

- There was no statistically significant difference in accuracy between the LSTM and GRU models ($p = 0.9249$).
- Both the LSTM and GRU models significantly outperformed the Transformer model in terms of classification accuracy ($p < 0.01$).

Table 4. Results of Statistical Significance Tests (Paired *t*-test).

Comparison	<i>p</i> -Value	Interpretation
LSTM vs. GRU	0.9249	Difference is not statistically significant
LSTM vs. Transformer	0.0066	LSTM performs significantly better
GRU vs. Transformer	0.0061	GRU performs significantly better

Across ten randomized runs on the dual-camera pose dataset (103 participants; 2060 labeled sequences), both recurrent attention-based models achieved high and stable performance. LSTM + Attention and GRU + Attention yielded mean accuracies of $\approx 98.9\%$ with narrow confidence bounds and low run-to-run variance; paired tests indi-

cated no significant difference between them ($p = 0.9249$). In contrast, the Transformer baseline averaged $\approx 96.6\%$ with visibly higher dispersion across seeds. Class-wise ROC analyses were uniformly high (AUCs approaching 1.0) for all models, yet the Transformer exhibited larger fluctuations across folds. Confusion matrices revealed that the principal residual errors concentrated on visually similar exercises—most notably “Biceps Curl” vs. “Lateral Raise”—with the misclassification rate for this pair highest under the Transformer. Attention-weight inspections on misclassified sequences indicated that both recurrent models concentrated salience around phase transitions (lift–lower boundaries), aligning with their lower variance and sharper decision boundaries.

To further evaluate the robustness of the models, we analyzed classification accuracy across three exercise difficulty levels: Low, Medium, and High. Figure 16 illustrates that all three architectures (LSTM + Attention, GRU + Attention, and Transformer) maintain consistently high accuracy across difficulty categories, indicating that the models generalize well regardless of movement complexity.

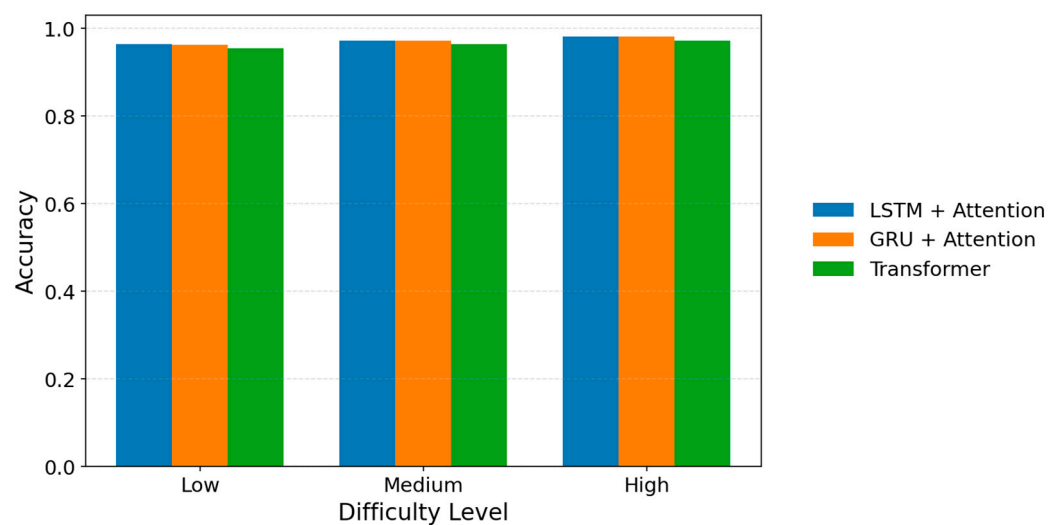


Figure 16. Classification accuracy of the three deep learning models across exercise difficulty levels (Low, Medium, High).

Despite slight variations—where the Transformer shows marginally lower performance in all three categories—the accuracy differences remain minimal (≤ 0.01), demonstrating stable recognition even as exercises become more demanding. This consistency suggests that the temporal–spatial representations extracted from dual-camera pose sequences are sufficiently discriminative across varying levels of biomechanical challenge.

Exercise duration is an important temporal factor that may influence pose stability, movement consistency, and model discriminative performance. Longer sequences can provide richer temporal information for recurrent architectures, whereas shorter sequences may contain more variability due to rapid transitions in posture. To examine how duration affects classification accuracy, the dataset was grouped into five time intervals: 0–10 s, 10–20 s, 20–30 s, 30–40 s, and 40–50 s.

Figure 17 presents the accuracy trends of LSTM + Attention, GRU + Attention, and Transformer models across these duration ranges.

As shown in Figure 17, both LSTM + Attention and GRU + Attention models exhibit a clear upward trend in accuracy as exercise duration increases from short (0–10 s) to moderate (20–30 s) intervals. This suggests that longer temporal windows provide more stable pose patterns, which are particularly beneficial for recurrent architectures. Accuracy slightly decreases in the 40–50 s interval, likely due to fatigue-related inconsistencies or increased intra-class variation.

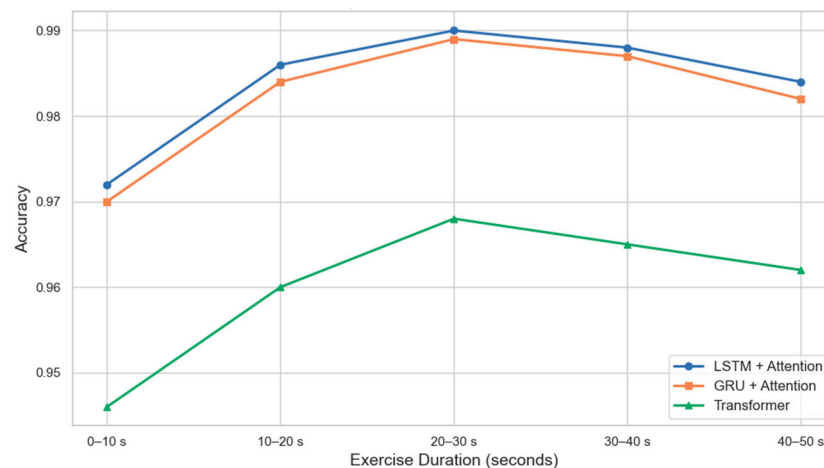


Figure 17. Model Accuracy Across Different Exercise Durations.

In contrast, the Transformer model demonstrates a more modest improvement with duration, showing limited gains beyond the 10–20 s interval. This indicates that, within this dataset, self-attention alone may be less effective than recurrent structures in leveraging fine-grained temporal dependencies for short exercise sequences. Overall, the duration-based analysis highlights how temporal characteristics of exercise execution interact with different model architectures.

These findings suggest that, within the context of the present dataset and the exercise-recognition task, the Transformer-based architecture is less effective relative to attention-augmented recurrent baselines. In contrast, RNN-based models—particularly when enhanced with attention—demonstrate more robust and reliable performance, combining superior mean accuracy with greater stability across random initializations and folds. Taken together with the confusion-structure and ROC evidence, the results indicate that inductive biases favoring temporal continuity and phase localization confer a measurable advantage for short, structured human-movement sequences.

5. Discussion

A close inspection of the error patterns shows that, despite the diagonal dominance of the confusion matrices, cross-confusions—most notably between “Biceps Curl” and “Lateral Raise”—remain more pronounced for the Transformer (Figures 10–15). Although the individual components (MediaPipe pose extraction, LSTM/GRU/Transformer architectures) are well-established, the present work provides a methodological advance by integrating (i) synchronized dual-camera temporal acquisition, (ii) multimodal fusion of anthropometrics with pose sequences, and (iii) statistically rigorous 10-run model comparison. These elements collectively provide new insights into viewpoint robustness, personalization, and model stability—areas that remain insufficiently explored in existing exercise-recognition literature. This suggests that local similarities in joint-angle configurations can blur class boundaries when temporal context is not modeled with sufficiently strong inductive biases. In exercise classification tasks where fine-grained temporal cues are critical, recurrent memory dynamics, especially when coupled with attention, appear to yield more separable decision regions than self-attention alone [37,42]. While class-wise ROC curves approach unity across all models, the comparatively greater fluctuation observed for the Transformer implies that, under limited and relatively homogeneous data regimes, self-attention may struggle to sustain stable discrimination at the class level [39,41]. These observations reinforce the benefit of mechanisms that prioritize “critical instants” along the sequence—e.g., frame-level or segment-level attention—to better capture phase transitions integral to movement identity.

Attention-map examinations on misclassified or low-confidence sequences further indicate that focus often shifts around repetition thresholds (e.g., lift–lower transitions). In the recurrent models, attention weights concentrate more tightly on these transition phases, which enhances signal-to-noise separation; by contrast, the Transformer’s multi-head focus tends to be broader, effectively “diluting” context over short windows. Methodologically, this motivates (i) sequence-length sensitivity analyses (alternative window sizes) and (ii) targeted data augmentation—temporal jitter, speed scaling, and viewpoint/scale perturbations—to sharpen decision boundaries. Consistent with Tables 3 and 4, the tightly clustered mean accuracies of LSTM + Attention and GRU + Attention ($\approx 98.9\%$) and their low variance indicate more reliable responses to phase changes, whereas the slightly higher standard deviation for GRU points to heightened sensitivity to weight initialization and optimization dynamics.

From an explainability standpoint, projecting attention weights onto the time axis highlights “instructional” frames (e.g., lockout in the shoulder press, heel-rise onset in the calf raise) that help disentangle visually similar postures. Because attention peaks are narrower and more concentrated for LSTM/GRU, these models may also be advantageous for real-time deployment under tight latency budgets: inference can allocate more compute to brief micro-windows containing decisive evidence while maintaining throughput at 30 FPS in a dual-camera setup [37,42]. This property aligns with the study’s practical aim of enabling real-time feedback without expert supervision.

In terms of generalizability, the dataset’s scope—103 participants performing five exercises—underscores the need for broader scene diversity (tempo variation, partial repetitions, grip width), clothing/illumination differences, and device heterogeneity to robustly assess out-of-distribution resilience. The Transformer’s higher variance plausibly reflects underexposure to such diversity; this variability may diminish with larger and more heterogeneous corpora and, crucially, with multimodal inputs (e.g., depth, IMU), which can supply complementary cues (perspective, micro-kinematics) that are muted in 2D pose trajectories [39,41]. In the interim, hybrid designs that combine recurrent temporal encoders with lightweight self-attention layers offer a pragmatic pathway, potentially harnessing the strengths of both paradigms until data scale and diversity catch up.

Personalization remains central to real-world performance. The demonstrated effects of BMI and athletic experience on accuracy motivate systematic inclusion of contextual and anthropometric covariates in the model input. In line with emerging work on context-aware systems in sports and rehabilitation [23,25], a multi-branch fusion layer that ingests demographic/anthropometric features (e.g., limb-segment ratios, flexibility proxies) alongside pose sequences may improve separation between look-alike movements and reduce the shift of errors from inter-class to inter-user variability. Such personalization could, in practice, yield more equitable performance across diverse populations and skill levels. Participants with higher BMI and lower training experience exhibited slightly increased misclassification rates, indicating that anthropometric variability may influence movement execution patterns and consequently affect model performance. This supports the importance of incorporating demographic covariates into personalized feedback systems.

For field deployment, several engineering considerations are salient: (i) complementing software-based dual-camera synchronization with periodic drift calibration or low-latency hardware triggers; (ii) augmenting viewpoint and illumination diversity to harden models against real-world noise; (iii) model compression (quantization; low-bit GEMMs) and pipeline parallelism to sustain edge-device throughput; and (iv) surfacing explanation artifacts in the user interface so that misclassifications come with human-interpretable evidence (“why this feedback now?”). These system-level refinements are well aligned with the stable recurrent performance reported in Tables 3 and 4 ($p = 0.9249$ for LSTM

vs. GRU; both significantly better than the Transformer) and would facilitate translating lab-grade results to everyday conditions.

Finally, ethical and trust considerations—privacy, informed consent, and transparency—will shape adoption in clinical and home settings. The study’s use of anonymized collection and written consent provides a sound basis for larger-scale trials; extending this with user-visible versioning/audit trails and rationale displays for exercise-level scores can accelerate trust formation and responsible use. Transitioning from these observations, the concluding paragraph’s emphasis on dataset expansion, in-the-wild evaluation, and integration of explainable AI modules follows naturally and sets a concrete agenda for scaling accessibility and confidence in autonomous feedback systems.

6. Conclusions

This study conducted a rigorous comparison of three temporal architectures—LSTM + Attention, GRU + Attention, and a Transformer baseline—for autonomous exercise recognition from dual-camera 2D pose trajectories in a cohort of 103 participants (2060 labeled samples). Across ten randomized runs, both recurrent attention-based models achieved high and stable accuracy (mean $\approx 98.9\%$) and significantly outperformed the Transformer (mean $\approx 96.6\%$). Paired *t*-tests confirmed no meaningful difference between LSTM and GRU ($p = 0.9249$), while both were superior to the Transformer ($p < 0.01$), a pattern consistent with confusion-matrix and ROC analyses indicating tighter class separation and lower variance for the recurrent models. These findings highlight that, in modest-scale, relatively homogeneous pose datasets, inductive biases for sequential continuity afforded by RNNs—amplified through attention—remain advantageous over purely self-attentional models. The analyses further showed that participant-specific covariates (e.g., BMI, athletic level) influence recognition performance, supporting the practical value of contextual features for personalization. Together, the results demonstrate a feasible, real-time pathway for expert-free exercise feedback with prospective utility in fitness, sports analytics, and rehabilitation settings.

Beyond these empirical outcomes, the work contributes: (i) a reproducible evaluation protocol with multi-run statistics and significance testing, (ii) an end-to-end pipeline for synchronized dual-camera capture and pose-based temporal modeling, and (iii) evidence that lightweight attention atop recurrent encoders can deliver both accuracy and stability for short, structured exercise sequences. Methodologically, our error analyses (confusion structure and ROC fluctuation) indicate that misclassifications concentrate around phase transitions for visually similar movements (e.g., biceps curl vs. lateral raise), underscoring the utility of attention to emphasize “critical instants” and motivating targeted augmentation (temporal jitter, speed scaling, viewpoint/illumination variation) to harden decision boundaries in real-world deployments.

Nevertheless, several limitations and opportunities for impact remain. The exercise taxonomy is deliberately constrained, and recordings were acquired in controlled conditions; broader scene diversity (tempos, partial repetitions, occlusions, clothing and lighting variations, device heterogeneity) and evaluation “in the wild” are needed to stress-test generalization. The reliance on 2D pose also suppresses depth and fine-grained kinematic cues; richer sensing promises to reduce ambiguity in look-alike patterns. Finally, while demographic features improved contextuality, systematic strategies for equitable personalization and calibration across subgroups are essential for inclusive performance.

Building on these results, several directions can be advanced to strengthen movement recognition and real-time utility. First, recognition robustness can be improved by fusing complementary modalities (RGB-D depth and IMU signals) with 2D pose and by exploring hybrid temporal encoders that combine recurrent filters for local continuity

with lightweight self-attention for longer-range dependencies. Second, model resilience under distribution shift can be advanced through targeted augmentation curricula (view-point/occlusion, lighting/clothing variation, tempo and partial-rep diversity), domain adaptation, and self-/semi-supervised objectives that leverage large unlabeled repetition corpora. Third, personalization and fairness can be enhanced via multi-branch fusion of anthropometric/demographic covariates with sequence embeddings, together with meta-learning or test-time adaptation to enable rapid per-user calibration and subgroup-aware evaluation to mitigate disparate error rates. Fourth, explainability and safety can be progressed by integrating attention/attribution overlays and calibrated uncertainty estimates directly into the interface, enabling human-interpretable evidence at the moment of decision. Crucially, a quantitative scoring framework can be incorporated to rate movement correctness (e.g., phase-specific form scores, range-of-motion completeness, tempo adherence), which can in turn be fed back into the model as supervision to further reinforce recognition and reduce confusions between visually similar exercises. In parallel, synchronous, real-time feedback can be strengthened by co-designing the end-to-end stack—pose estimation, temporal inference, and rendering—for edge devices using quantization, operator fusion, and low-bit GEMMs to meet tight latency budgets without compromising accuracy; closed-loop evaluations can then assess whether instantaneous cues improve technique and adherence. Finally, longitudinal, IRB-approved trials and the release of reproducible benchmarks, model cards, and privacy-preserving data schemas can extend the work's external validity and practical impact.

This work shows that attention-enhanced recurrent architectures provide a strong and stable foundation for autonomous exercise recognition in compact pose datasets, clarifies when and why they outperform Transformers, and outlines a practical route to multimodal, explainable, and equitable AI coaching at scale. By advancing robustness, personalization, quantitative scoring of technique, and truly synchronous feedback, we aim to translate lab-grade performance into trustworthy, everyday tools for athletes, patients, and practitioners.

Author Contributions: Conceptualization, U.Ö. and Ö.C.T.; methodology, U.Ö.; software, U.Ö.; validation, U.Ö. and Ö.C.T.; formal analysis, U.Ö.; investigation, U.Ö.; data curation, U.Ö.; writing—original draft preparation, U.Ö.; writing—review and editing, Ö.C.T.; visualization, U.Ö.; supervision, Ö.C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Formal ethical approval was not sought for this study as it was deemed to involve minimal risk and utilized anonymous data collection procedures. Written informed consent was obtained from all individual participants involved in the study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Each participant provided written consent confirming voluntary participation and permission for anonymized data processing.

Data Availability Statement: The data that support the findings of this study are not publicly available due to ongoing use in another study but are available from the corresponding author upon reasonable request.

Acknowledgments: This study was conducted as part of the first author's Ph.D. thesis at İstanbul University—Cerrahpaşa, Department of Computer Engineering. The authors would like to thank all participants who contributed to the data collection process. During the preparation of this manuscript, the authors used ChatGPT (GPT-5, OpenAI) for language editing and formatting suggestions. The authors have reviewed and edited the output and take full responsibility for the final content.

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

References

1. Hammes, F.; Hagg, A.; Asteroth, A.; Link, D. Artificial intelligence in elite sports—A narrative review of success stories and challenges. *Front. Sports Act. Living* **2022**, *4*, 861466. [\[CrossRef\]](#)
2. Gill, K.S.; Anand, V.; Malhotra, S.; Devliyal, S. Sports Game Classification and Detection Using ResNet50 Model Through Machine Learning Techniques Using Artificial Intelligence. In Proceedings of the 2024 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India, 1–3 March 2024; pp. 1–5.
3. Gajendra, V. Reliability of Expert Judgement Elicitation in the field of Multi-Criteria Decision Analysis. Master's Thesis, College of Engineering, Western New England University, Springfield, MA, USA, 2024.
4. Mohammed, A.H.; Othman, Z.J.; Abdullah, A.I. The Role of Artificial Intelligence in Enhancing Sports Analytics and Training. *Cihan Univ. Erbil Sci. J.* **2024**, *8*, 58–62. [\[CrossRef\]](#)
5. Novatchkov, H.; Baca, A. Artificial intelligence in sports on the example of weight training. *J. Sports Sci. Med.* **2013**, *12*, 27.
6. Ratiu, O.G.; Badau, D.; Carstea, C.G.; Badau, A.; Paraschiv, F. Artificial intelligence (AI) in sports. In Proceedings of the 9th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Cambridge, UK, 20–22 February 2010; pp. 93–97.
7. Rodrigues, G.; Madiwale, M.N. Evolution of Sports with Artificial Intelligence. *ITM Web Conf.* **2024**, *68*, 01001. [\[CrossRef\]](#)
8. Zhang, Y.; Li, H.; Huang, R. The Effect of Tai Chi (Bafa Wubu) Training and Artificial Intelligence-Based Movement-Precision Feedback on the Mental and Physical Outcomes of Elderly. *Sensors* **2024**, *24*, 6485. [\[CrossRef\]](#)
9. Annesi, J.J. Effects of computer feedback on adherence to exercise. *Percept. Mot. Ski.* **1998**, *87*, 723–730. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Kotte, H.; Daiber, F.; Kravcik, M.; Duong-Trung, N. Fitsight: Tracking and feedback engine for personalized fitness training. In Proceedings of the 32nd ACM Conference User Modeling, Adaptation and Personalization 2024, Sardinia, Italy, 1–4 July 2024; pp. 223–231.
11. Bays, D.K.; Verble, C.; Verble, K.M.P. A Brief Review of the Efficacy in Artificial Intelligence and Chatbot-Generated Personalized Fitness Regimens. *Strength Cond. J.* **2024**, *46*, 485–492. [\[CrossRef\]](#)
12. Maulana, M.F.; Okamoto, Y.; Sato-Shimokawara, E. Creating Feedback to Maintain Motivation when Doing Home Based Exercise Using Pose Estimation: Single Subject Design. In Proceedings of the 2024 International Electronics Symposium (IES), Bali, Indonesia, 6–8 August 2024; pp. 335–340.
13. Zhu, J.; Dallal, D.H.; Gray, R.C.; Villareale, J.; Ontañón, S.; Forman, E.M.; Arigo, D. Personalization paradox in behavior change apps: Lessons from a social comparison-based personalized app for physical activity. In Proceedings of the ACM Human-Computer Interaction, Virtual, 23–27 October 2021; Volume 5, pp. 1–21.
14. Kuru, H. Identifying behavior change techniques in an artificial intelligence-based fitness app: A content analysis. *Health Educ. Behav.* **2024**, *51*, 636–647. [\[CrossRef\]](#)
15. Kotte, H.; Kravcik, M.; Duong-Trung, N. Real-Time Posture Correction in Gym Exercises: A Computer Vision-Based Approach for Performance Analysis. Error Classification and Feedback. In Proceedings of the MILEs@ EC-TEL, Aveiro, Portugal, 4–8 September 2023; pp. 64–70.
16. Dhakate, H.; Anasane, S.; Shah, S.; Thakare, R.; Rawat, S.G. Enhancing Yoga Practice: Real-time Pose Analysis and Personalized Feedback. In Proceedings of the 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, 9–10 February 2024; pp. 35–40.
17. Sitapara, S.; Akarsha, D.P.; Jain, M.; Singh, S.; Gupta, N. Novel Approach for Real-Time Exercise Posture Correction Using Computer Vision and CNN. In Proceedings of the 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIE), Bellary, India, 2–3 November 2023; pp. 1–6.
18. Alatiyah, T.; Chen, C. Recognizing exercises and counting repetitions in real time. *arXiv* **2020**, arXiv:2005.03194. [\[CrossRef\]](#)
19. Fieraru, M.; Zafir, M.; Pirlea, S.C.; Olaru, V.; Sminchisescu, C. Aifit: Automatic 3D human-interpretable feedback models for fitness training. In Proceedings of the IEEE/CVF Conf. Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 9919–9928.
20. Duong-Trung, N.; Kotte, H.; Kravcik, M. Augmented intelligence in tutoring systems: A case study in real-time pose tracking to enhance the self-learning of fitness exercises. In Proceedings of the European Conference on Technology Enhanced Learning, Aveiro, Portugal, 4–8 September 2023; pp. 705–710.
21. Nagarkoti, A.; Teotia, R.; Mahale, A.K.; Das, P.K. Realtime indoor workout analysis using machine learning & computer vision. In Proceedings of the 2019 41st Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 1440–1443.
22. Zhang, W.; Chen, X. Theoretical Application Analysis of Deep Learning Algorithms in Sports Injury Prediction. In Proceedings of the 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), Marseille, France, 23–25 September 2024; pp. 886–891.
23. Mekruksavanich, S.; Phaphan, W.; Hnoohom, N.; Jitpattanakul, A. Recognition of sports and daily activities through deep learning and convolutional block attention. *PeerJ Comput. Sci.* **2024**, *10*, e2100. [\[CrossRef\]](#)

24. Musat, C.L.; Mereuta, C.; Nechita, A.; Tutunaru, D.; Voipan, A.E.; Voipan, D.; Mereuta, E.; Gurau, T.V.; Gurău, G.; Nechita, L.C. Diagnostic Applications of AI in Sports: A Comprehensive Review of Injury Risk Prediction Methods. *Diagnostics* **2024**, *14*, 2516. [\[CrossRef\]](#)
25. Guan, L. Intelligent rehabilitation assistant: Application of deep learning methods in sports injury recovery. *Mol. Cell. Biomech.* **2024**, *21*, 384. [\[CrossRef\]](#)
26. Tang, D. Hybridized hierarchical deep convolutional neural network for sports rehabilitation exercises. *IEEE Access* **2020**, *8*, 118969–118977. [\[CrossRef\]](#)
27. Cust, E.E.; Sweeting, A.J.; Ball, K.; Robertson, S. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *J. Sports Sci.* **2019**, *37*, 568–600. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Liao, Y.; Vakanski, A.; Xian, M. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 468–477. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Johnson, W.R.; Mian, A.; Lloyd, D.G.; Alderson, J.A. On-field player workload exposure and knee injury risk monitoring via deep learning. *J. Biomech.* **2019**, *93*, 185–193. [\[CrossRef\]](#)
30. Lu, L. Intelligent Training Assistance System for Sports Posture Correction Feedback Based on Artificial Intelligence Algorithms. In Proceedings of the 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT), Bangalore, India, 15–16 March 2024; pp. 1–6.
31. Annino, G.; Bonaiuto, V.; Campoli, F.; Caprioli, L.; Edriss, S.; Padua, E.; Panichi, E.; Romagnoli, C.; Romagnoli, N.; Zanela, A. Assessing sports performances using an artificial intelligence-driven system. In Proceedings of the 2023 IEEE International Workshop on Sport, Technology and Research (STAR), Trento, Italy, 14–16 September 2023; pp. 98–103.
32. Krstić, D.; Vučković, T.; Dakić, D.; Ristić, S.; Stefanović, D. The application and impact of artificial intelligence on sports performance improvement: A systematic literature review. In Proceedings of the 2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), Plovdiv, Bulgaria, 23–25 November 2023; pp. 1–8. [\[CrossRef\]](#)
33. Künzell, S.; Knoblich, A.; Stippler, A. Valid knowledge of performance provided by a motion capturing system in shot put. *Front. Sports Act. Living* **2025**, *6*, 1482701. [\[CrossRef\]](#)
34. Chidambaram, S.; Maheswaran, Y.; Patel, K.; Sounderajah, V.; Hashimoto, D.A.; Seastedt, K.P.; McGregor, A.H.; Markar, S.R.; Darzi, A. Using artificial intelligence-enhanced sensing and wearable technology in sports medicine and performance optimisation. *Sensors* **2022**, *22*, 6920. [\[CrossRef\]](#)
35. Parisi, G.I.; von Stosch, F.; Magg, S.; Wermter, S. Learning human motion feedback with neural self-organization. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–6.
36. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Grundmann, M. Mediapipe: A framework for perceiving and processing reality. In Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
37. Zhong, G.; Lin, X.; Chen, K.; Li, Q.; Huang, K. Long short-term attention. In Proceedings of the Advances in Brain Inspired Cognitive Systems: 10th International Conference, BICS 2019, Guangzhou, China, 13–14 July 2019; Proceedings 10. Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 45–54.
38. Bhuvaneswari, M.; Praba, V.L. An Illustration of Attention Mechanism in LSTM Algorithm for Analyzing Its Performance. *Int. J. Health Sci.* **2022**, *6*, 9010–9023. [\[CrossRef\]](#)
39. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks. *arXiv* **2023**, arXiv:2306.07303. [\[CrossRef\]](#)
40. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Karita, S.; Chen, N.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N.E.Y.; Yamamoto, R.; Wang, X.; et al. A comparative study on transformer vs rnn in speech applications. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 449–456. [\[CrossRef\]](#)
42. Mohsen, S. Recognition of human activity using GRU deep learning algorithm. *Multimed. Tools Appl.* **2023**, *82*, 47733–47749. [\[CrossRef\]](#)
43. Bai, C. AGA-GRU: An optimized GRU neural network model based on adaptive genetic algorithm. *J. Phys. Conf. Ser.* **2020**, *1651*, 012146. [\[CrossRef\]](#)
44. Lin, Z. Enhanced GRU-based regression analysis via a diverse strategies whale optimization algorithm. *Sci. Rep.* **2024**, *14*, 25629. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.