

ALEX-GYM-1: A Novel Dataset and Hybrid 3D Pose Vision Model for Automated Exercise Evaluation

Ahmed Hassan¹^{a*}, Abdelaziz Serour¹^{b*}, Ahmed Gamea¹^{c*} and Walid Gomaa^{1,2}^d

¹Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt

²Faculty of Engineering, Alexandria University, Alexandria, Egypt

Keywords: Multi-Modal Deep Learning, Exercise Assessment, Computer Vision, Pose Estimation, Human Movement Analysis, Fitness Monitoring.

Abstract: Improper gym exercise execution often leads to injuries and suboptimal training outcomes, yet conventional assessment relies on subjective human observation. This paper introduces ALEX-GYM-1, a novel multi-camera view dataset with criterion-specific annotations for squats, lunges, and Romanian deadlifts, alongside a complementary multi-modal architecture for automated assessment. Our approach uniquely integrates: (1) a vision-based pathway using 3D CNN to capture spatio-temporal dynamics from video, and (2) a pose-based pathway that analyzes biomechanical relationships through engineered landmark features. Extensive experiments demonstrate the superiority of our Multi-Modal fusion architecture over both single-modality methods and competing approaches, achieving Hamming Loss reductions of 30.0% compared to Vision-based and 79.5% compared to Pose-based models. Feature-specific analysis reveals key complementary strengths, with Vision-based components excelling at contextual assessment (89% error reduction for back knee positioning) while Pose-based components demonstrate precision in specific joint relationships. The computational efficiency analysis enables practical deployment strategies for both real-time edge applications and high-accuracy cloud computing scenarios. This work addresses critical gaps in exercise assessment technology through a purpose-built dataset and architecture that establishes a new state-of-the-art for automated exercise evaluation in multi-camera view settings.

1 INTRODUCTION

Improper execution of resistance exercises represents a significant public health concern, with epidemiological studies revealing alarming injury patterns among strength athletes. Research by (Winwood et al., 2014) documented that 82% of strongman competitors experienced training-related injuries, with the lower back (24%) and knees (11%) being the most commonly affected anatomical locations. Similarly, (Wushao et al., 2000) reported high incidence rates among weightlifters, with lower back injuries accounting for 19% in male athletes and 18% in female athletes, while knee injuries represented 29% in males and 32% in females. These statistics un-

derscore the critical need for improved exercise form monitoring, as injuries not only disrupt training progression but often lead to chronic conditions requiring extensive rehabilitation (Keogh and Winwood, 2017). While professional trainers provide personalized feedback, this traditional approach is limited by subjectivity, variability across trainers, and accessibility challenges, especially in remote or home-based training environments (Clark et al., 2021).

Recent technological solutions have attempted to address these limitations through automated exercise assessment systems. Kotte et al. (Kotte et al., 2023) introduced a computer vision approach using YOLOv7-pose for keypoint detection, offering real-time posture correction feedback through comparison with expert demonstrations. Similarly, Duong-Trung et al. (Duong-Trung et al., 2023) developed a pose tracking system that learns from either live professional trainer demonstrations or recorded expert videos to provide automated guidance. Neha and

^a <https://orcid.org/0009-0000-9177-3815>

^b <https://orcid.org/0009-0007-9601-9759>

^c <https://orcid.org/0009-0004-6023-4420>

^d <https://orcid.org/0000-0002-8518-8908>

* These authors contributed equally to this work.

Manju (Neha and Manju, 2023) utilized MediaPipe with OpenCV to implement a virtual fitness trainer that evaluates form and technique through computer vision. VCOACH (Youssef et al., 2023) presented a more sophisticated approach using distance matrices derived from MediaPipe pose estimations as input to a residual neural network for simultaneous exercise recognition and error assessment. However, these solutions face critical limitations: they predominantly rely on binary correctness assessment rather than detailed biomechanical criteria, depend heavily on exact reference poses from experts that may not accommodate individual body proportions, lack fine-grained analysis of exercise-specific biomechanical principles, or utilize single-modality inputs that miss complementary information available in direct visual analysis. Additionally, most existing systems rely exclusively on either raw visual data or skeletal landmarks, missing the complementary benefits of integrating both modalities for comprehensive assessment.

In this research, exercise assessment is formulated as a multi-label binary classification task for predetermined exercise types. For each exercise—squat, lunge, or Romanian deadlift—5-7 biomechanical criteria identified by professional trainers are evaluated simultaneously, enabling specific identification of form deficiencies rather than providing simplistic binary judgments. A dual-stream architecture processes both visual data through 3D convolutional neural networks and biomechanical relationships through engineered pose features. Visual streams capture contextual movement qualities and temporal dynamics, while pose streams quantify precise joint relationships, with these complementary information sources integrated through a multi-stage fusion network. This approach transcends the limitations of single-modality systems by enabling assessment capabilities unachievable through either visual or skeletal analysis alone.

Leveraging this approach, our study introduces ALEX-GYM-1, a purpose-built dataset for exercise analysis, alongside a novel hybrid assessment system designed to address the identified limitations. The primary contributions of this work are as follows:

- **Novel Multi-View Dataset:** Development of ALEX-GYM-1, featuring synchronized frontal and lateral recordings with criterion-specific annotations across fundamental training exercises, providing the comprehensive reference data necessary for detailed biomechanical assessment.
- **Multi-Modal Architecture:** Design of a hybrid framework that integrates complementary visual and biomechanical information streams, over-

coming the inherent limitations of single-modality approaches through structured fusion of contextual and skeletal data.

- **Biomechanical Criteria-Based Model Comparison:** Systematic analysis of how vision-based, pose-based, and multi-modal approaches perform across specific biomechanical assessment criteria.
- **Edge Deployment Performance Evaluation:** Implementation and testing of the proposed models on edge computing hardware to determine inference latency, enabling informed deployment decisions for real-time exercise assessment applications.

This system advances real-time exercise assessment, providing objective and detailed feedback crucial for injury prevention and performance optimization.

The paper is organized as follows. Section 1 is an introduction. Section 2 reviews related literature, Section 3 details our methodology, Section 4 presents results and analysis, and Section 5 concludes with key findings and future directions.

2 RELATED WORK

2.1 Human Activity Recognition and Exercise Assessment

Human Activity Recognition (HAR) has evolved through two primary approaches: wearable sensor-based and vision-based modalities (Kaseris et al., 2024). Vision-based methods predominantly rely on visible-light and depth cameras, with much of the research focusing on either raw image inputs or skeletal pose representations (Baradel et al., 2018; Khan et al., 2022). Building on this foundation, recent research has increasingly focused on exercise evaluation using vision-based inputs, particularly skeletal pose data and RGB frames extracted from video. This trend is driven by the need for automated fitness and rehabilitation systems that not only recognize activity type but also assess execution quality.

Ganesh et al. (Ganesh et al., 2020) developed a virtual coaching system using RGB video and OpenPose (Cao et al., 2019) for 2D keypoint extraction, achieving 98.98% accuracy in exercise recognition with a Random Forest classifier. Their assessment method compared user and trainer skeletons by averaging point-wise distances, but this approach failed to distinguish genuine mistakes from variations in body proportions, offering limited and inaccurate feedback.

In VCOACH (Youssef et al., 2023), the authors proposed an efficient architecture for simultaneous exercise recognition and detailed performance evaluation, leveraging 1D convolutional layers with residual blocks applied to skeletal pose keypoint sequences extracted via MediaPipe pose estimation. The system was designed to handle multi-label classification in addition to identifying the type of exercise. Evaluated on the MMDOS dataset (Sharshar et al., 2022), the model achieved an average accuracy of $75.34\% \pm 6.71\%$ and a Jaccard score of 0.71 ± 0.03 across ten runs under the Full Assessment Configuration, which targets the detection of specific execution mistakes alongside activity recognition.

2.2 Vision-Based Models for Video Analysis

Recent advances in vision models have demonstrated strong capabilities in processing RGB image sequences for video-based understanding tasks. Pre-trained architectures such as Vision Transformers (ViT) (Dosovitskiy et al., 2020) and 3D ResNet (Du et al., 2023) have shown significant success across a range of applications due to their ability to capture spatial and temporal dependencies. In our work, we adopt a set of vision-based architectures to process RGB video data for exercise recognition and assessment. These include 3D ResNet (Du et al., 2023), 1D CNN-GRU combined with ViT (Chen et al., 2023), and a 2D CNN + ViT hybrid (Arnab et al., 2021). The detailed architecture and implementation of each approach are discussed in later sections.

2.3 Datasets for Exercise Analysis

Most existing datasets focus on activity recognition rather than detailed correctness assessment. Table 1 compares key exercise analysis datasets, revealing significant gaps in multi-view recording, detailed correctness criteria, and exercise-specific biomechanical feature extraction.

General-purpose datasets like UTD-MHAD (Chen et al., 2015) (8 subjects, 27 actions) and MoVi (Ghorbani et al., 2021) (90 participants, 20 actions) offer technical comprehensiveness but lack correctness annotations. More specialized datasets include HSiPu2 (Zhang et al., 2021) with 3 exercises (push-up, pull-up, and sit-up) using dual-view recording but lacking participant count documentation, and KIMORE (Capecci et al., 2019) containing 78 participants (44 healthy, 34 with motor dysfunctions) performing 5 rehabilitation exercises with clinical scoring. The MEx dataset (Wijekoon

et al., 2019) captures 7 physiotherapy exercises performed by 30 subjects using multiple sensors (accelerometers, pressure mat, depth camera) but lacks exercise correctness annotations.

ALEX-GYM-1 addresses these limitations by uniquely combining:

- **Synchronized multi-camera view recording:** Frontal and lateral camera views capture complementary information - bilateral symmetry and balance from front, joint angles and postural alignment from side - enabling comprehensive assessment impossible with single perspectives.
- **Fine-grained biomechanical criteria:** Rather than binary correct/incorrect labels, 5-7 specific criteria per exercise enable targeted feedback on distinct movement components, identified by professional trainers as critical for performance and safety.
- **Invariant biomechanical features:** Skeletal landmarks are transformed into distance and angular relationships that remain consistent regardless of body proportions, camera position, or execution speed, creating robust representations of movement quality.

The novelty of the ALEX-GYM-1 dataset can be observed in its comprehensive integration of multi-view recordings with criterion-specific biomechanical annotations. Whereas existing datasets have been limited to either general activity recognition or simplified binary correctness labels, this dataset was purposefully designed to bridge the gap between mere exercise identification and detailed movement quality assessment. The dataset's synchronized dual-perspective recordings, combined with expert-validated biomechanical criteria and standardized feature representations, establish a new benchmark for exercise assessment resources. This structured approach enables more nuanced exercise evaluation than previously possible, allowing for the development of automated systems capable of providing specific, actionable feedback comparable to that of human trainers.

3 METHODOLOGY

3.1 Dataset

The ALEX-GYM-1 dataset was constructed to capture fundamental gym exercises performed by individuals of varying demographics. Videos of squats, lunges, and single-leg Romanian deadlifts were recorded from both frontal and lateral camera views,

Table 1: Comprehensive comparison of workout exercise analysis datasets.

Dataset	Multi-View	Pose Data	Correctness Criteria	Participants	Expert Annotation	Biomechanical Features
ALEX-GYM-1	Yes (2)	3D	Multi-label	45	Yes	Yes
HSiPu2 (Zhang et al., 2021)	Yes (2)	2D	Binary	Not reported	Yes	No
KIMORE (Capecci et al., 2019)	No	3D	Multi-score	78	Yes	Partial
UTD-MHAD (Chen et al., 2015)	No	3D	No	8	No	No
MEx (Wijekoon et al., 2019)	No	No	No	30	No	No
MoVi (Ghorbani et al., 2021)	Yes (multi)	MoCap	No	90	No	No

with professional trainers annotating specific biomechanical criteria—each represented as a boolean feature—to assess movement correctness.

3.1.1 Demographic Composition

The dataset comprises 45 participants distributed across demographic categories as presented in Table 2.

Table 2: Demographic distribution of participants in ALEX-GYM-1.

Gender	
Female	12
Male	33
Age Groups	
Children (8–12 years)	4
Teenagers (13–19 years)	7
Young Adults (20–35 years)	31
Adults (36–54 years)	3

3.1.2 Exercise Composition

The dataset contains 670 videos distributed across three exercise types:

- *Squat (295 videos)*: Six boolean features were annotated: feet flat, simultaneous hip/knee bend, backward hip movement, neutral lower back, hips below knees, feet angled 30°.
- *Lunges (106 videos)*: Seven boolean features were annotated: shoulder-width heels, forward gaze, 90° knee bend, natural arm movement, aligned feet, straight back, back knee positioning.
- *Single-Leg Romanian Deadlifts (269 videos)*: Five boolean features were annotated: balance maintenance, back alignment, full leg extension, controlled reversal, support knee angle.

All data collection was conducted with participants’ informed consent for research purposes.

3.2 Data Preprocessing

A two-stage preprocessing pipeline was implemented to standardize inputs and extract biomechanically-relevant features.

General preprocessing: Original videos were recorded at 30 FPS with durations ranging from 3 to 8 seconds, depending on exercise execution and participant speed. To normalize temporal variation, each video was temporally resampled by extracting 16 uniformly spaced frames per camera view (Frontal and Lateral) at intervals of video-duration/16, creating a standardized representation regardless of original performance speed. MediaPipe’s 3D pose estimation was applied to extract 33 skeletal landmarks per frame.

The selection of 16 frames was determined through quantitative analysis of the trade-off between computational efficiency and motion fidelity. A frame stacking methodology was employed, wherein binary silhouettes were extracted from frames and aggregated through pixel-wise summation to generate composite motion heatmaps. These temporal aggregations were evaluated at varying frame counts (4, 8, 16, 32) against a 64-frame reference standard using Structural Similarity Index (SSIM) and Intersection over Union (IoU) metrics.

Table 3: Comparative Analysis of Motion Representation Fidelity Across Frame Sampling Rates.

Frame Count	Frontal View (SSIM / IoU)	Lateral View (SSIM / IoU)
4	0.924 / 0.982	0.946 / 0.990
8	0.946 / 0.995	0.948 / 0.993
16	0.967 / 0.997	0.963 / 0.998
32	0.985 / 0.999	0.982 / 0.999

While utilizing higher frame counts approached perfect representation, computational overhead increased proportionally. The 16-frame configuration preserved over 96% SSIM and exceeded 99.7% IoU across both camera views, offering an optimal balance between computational efficiency and motion representation fidelity.

Pose feature engineering: While the extracted skeletal landmarks provide positional information, raw coordinates are highly sensitive to variations in subject positioning, camera angle, and body proportions. To create robust, invariant representations suitable for exercise assessment, these landmarks were transformed into biomechanically meaningful relationships through two complementary feature sets:

Euclidean distances. Euclidean distances between

landmark pairs were calculated:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

where $p_i = (x_i, y_i, z_i)$ and $p_j = (x_j, y_j, z_j)$ represent keypoint coordinates. Due to symmetry ($d_{ij} = d_{ji}$), the total unique distances was determined as

$$\text{Total Pairwise Distances} = \frac{33 \times 32}{2} = 528$$

Angular Relationships. Angular relationships between landmark triplets were derived, with cosine values used for numerical stability:

$$\begin{aligned} \theta_{ijk} &= \cos^{-1} \left(\frac{(p_i - p_j) \cdot (p_k - p_j)}{\|p_i - p_j\| \|p_k - p_j\|} \right) \quad (2) \\ \Rightarrow \cos(\theta_{ijk}) &= \frac{(p_i - p_j) \cdot (p_k - p_j)}{\|p_i - p_j\| \|p_k - p_j\|} \end{aligned}$$

where p_i , p_j , and p_k represent three distinct landmarks with p_j serving as the vertex point. The vectors $(p_i - p_j)$ and $(p_k - p_j)$ form the two sides of the angle being measured. The total unique angle combinations was determined as

$$\text{Total Pose Angles} = \binom{33}{3} = \frac{33 \times 32 \times 31}{3 \times 2 \times 1} = 5456$$

creating a final feature vector of 5984 dimensions (528 distances + 5456 angles) that provide a position-invariant representation of biomechanical relationships.

3.3 Data Organization

The dataset was structured to facilitate efficient access and processing:

- **Demographic metadata:** Excel file containing participant attributes including age, gender, height, and weight.
- **Technical data:** Excel file with video identifiers for each camera view (odd/even for frontal/lateral), repetition indices, exercise classifications, and binary correctness criteria for each biomechanical criteria.
- **Naming convention:** Files follow $(\text{num_video})_idx_(\text{repetition_idx})_(\text{frame_number}).jpg$ as shown in Figure 1.
- **Directory structure:** Three primary directories (Squats, Lunges, Deadlifts) with hierarchical organization by camera view and repetition.



Figure 1: Image naming convention demonstrating corresponding frontal (1_idx_0_1.jpg) and lateral (2_idx_0_1.jpg) camera views.

3.4 Model Architecture

To effectively classify these diverse biomechanical criteria, which range from precise joint angles to complex movement patterns, a multi-modal architecture capable of capturing both explicit skeletal relationships and visual contextual information was developed. This dual perspective design enables the extraction of complementary information from each modality, resulting in a more accurate assessment than using any single modality alone.

3.4.1 System Overview

In this study, exercise evaluation is formulated as a multi-label binary classification task. The proposed system does not classify the exercise type itself, as this is predetermined through dataset organization prior to assessment. Instead, for each known exercise, the system evaluates a set of specific biomechanical criteria (6 for squats, 5 for deadlifts, and 7 for lunges), where each criterion is represented as a binary value indicating whether it is satisfied (1) or not satisfied (0) in the exercise execution. This approach enables detailed, criterion-wise feedback on exercise form.

As illustrated in Figure 2, the proposed processing framework employs a dual-stream architecture that integrates both visual and skeletal modalities. The system processes 16 frames from each camera view, with all three RGB color channels preserved to maintain visual details. The system features two parallel pathways: a pose-based stream, where 33 landmarks extracted via MediaPipe are transformed into a 5984-dimensional vector for position-invariant representation, and a vision-based stream, where raw video frames are processed using specialized 3D convolutional networks to capture spatio-temporal dynamics. A key innovation of this architecture is its operational flexibility—each modality is capable of functioning independently as a standalone assessment system or collaboratively within a multi-modal fusion

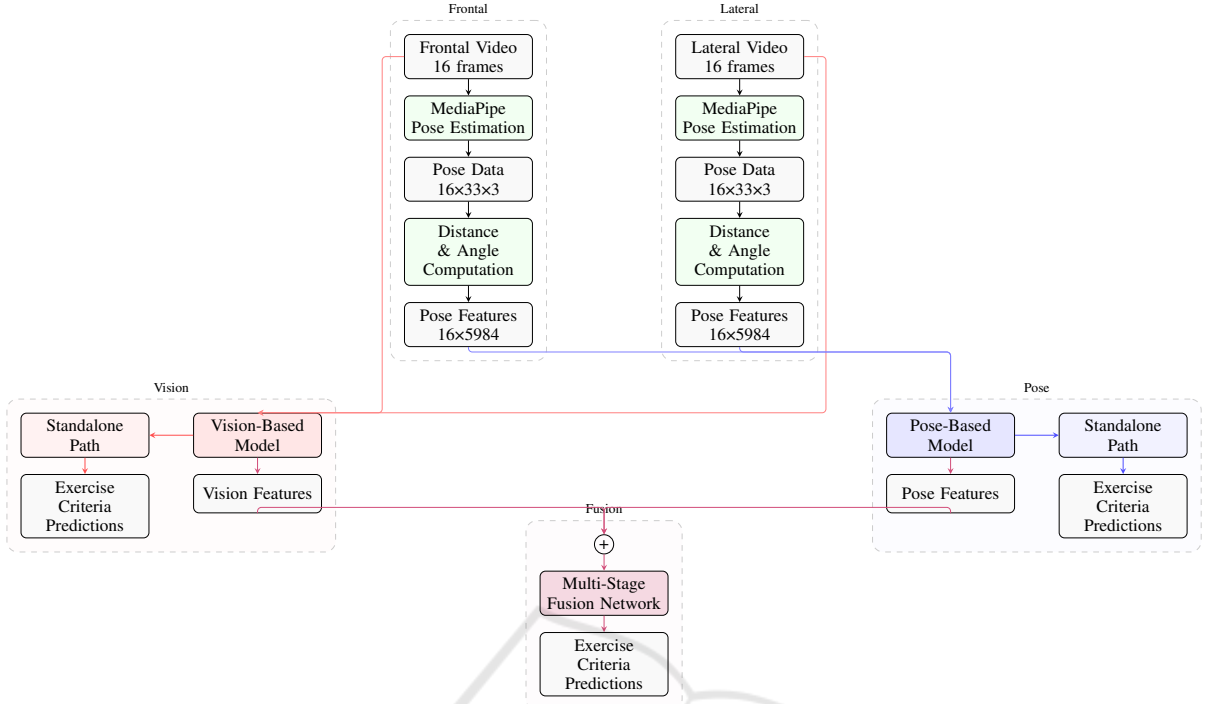


Figure 2: Multi-Modal Exercise Evaluation System workflow: Vision (red) and Pose (blue) pathways operate independently or through fusion.

framework. This modular design supports deployment under varying computational constraints, enabling a lightweight option in resource-limited settings while preserving the option for full multi-modal analysis when higher accuracy is required. The information flow is driven by synchronized frontal and lateral video inputs, offering multi-view perspectives that mitigate occlusion and ambiguity common in single-view setups. Each pathway is carefully optimized to leverage the strengths of its respective data modality.

3.4.2 Vision-Based Model

The Vision-Based Model, illustrated in Figure 3, employs a dual-stream 3D convolutional architecture designed to process frontal and lateral camera views separately before integrating their outputs for comprehensive analysis. Each camera view (Frontal and Lateral) is handled by a dedicated ResNet3D-18 network pre-trained on the Kinetics-400 action recognition dataset. A selective transfer learning strategy was adopted, where early layers (labeled as "Frozen") were kept static to preserve their ability to extract low-level visual features such as edges, textures, and basic motion cues, while deeper layers (labeled as "Fine-tuned") were adapted to specialize in exercise-specific movement patterns. This approach reduces training

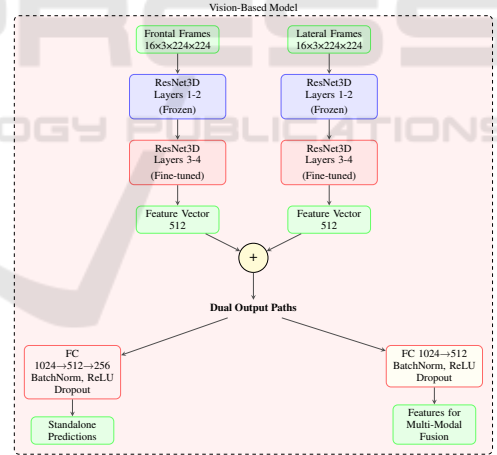


Figure 3: Vision-Based Model architecture with dual ResNet3D streams. Frontal and lateral frames are processed through transfer-learned 3D CNN pathways with selectively frozen and fine-tuned layers.

complexity, accelerates convergence, and improves generalization by balancing pre-learned visual representations with task-specific adaptation. The resulting feature vectors from each stream, capturing essential spatio-temporal characteristics, are concatenated into a unified representation. This fused representation is then directed either into a standalone model path or into a feature extraction path for integration

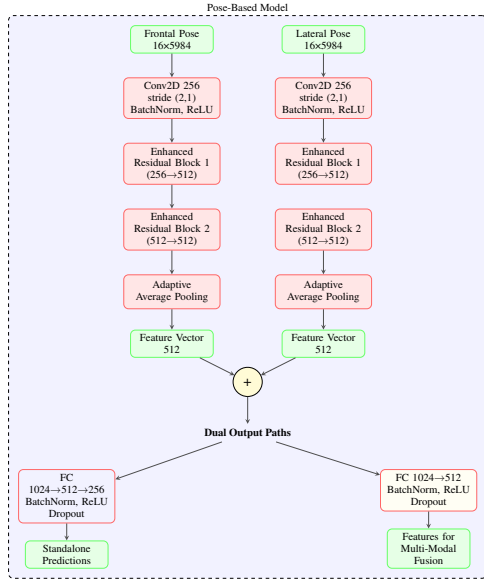


Figure 4: Enhanced Pose-Based Model architecture. High-dimensional pose features from both camera views are processed through specialized convolutional pathways with multi-stage residual blocks and global pooling.

with pose-based information in the multi-modal fusion framework. Throughout the architecture, batch normalization, ReLU activations, and dropout layers are applied to ensure stable training and robust performance.

3.4.3 Pose-Based Model

The Pose-Based Model, illustrated in Figure 4, was designed to capture the biomechanical relationships between body joints. Unlike the Vision-Based Model, this component focuses exclusively on a structured representation of body configuration. The input to the model consists of high-dimensional vectors encoding pairwise distances and angular relationships between joints crafted to be invariant to position, scale, and partial rotation. To effectively process these high-dimensional features, the model first applies a 2D convolutional layer that reduces dimensionality while preserving the temporal structure of the sequence. The network uses residual blocks (shown in Figure 5) that include three main operations: channel reduction, feature processing, and channel expansion. These blocks incorporate shortcut connections that help information flow through deep layers, improving training stability. Following this processing, adaptive average pooling creates compact feature vectors for each camera view by consolidating temporal information. These representations are then concatenated and routed through either a standalone prediction pathway or a feature extraction pathway designed for integra-

tion with the visual stream in the multi-modal fusion framework, mirroring the architecture of the Vision-Based Model.

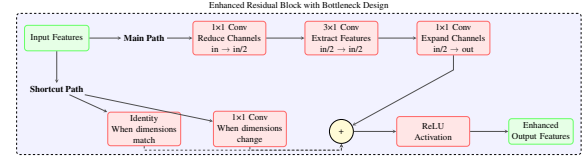


Figure 5: Residual Block with Bottleneck Design. This architecture implements a computationally efficient main path with three specialized convolutions (reduction→extraction→expansion), complemented by an adaptive shortcut for maintaining feature integrity, optimizing both gradient flow and inference efficiency.

3.4.4 Multi-Modal Feature Fusion

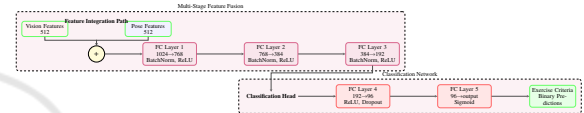


Figure 6: Multi-Modal Feature Fusion Architecture. Complementary feature sets are progressively combined and refined through fully connected layers with normalization and dimension reduction, culminating in a specialized classification head that produces criterion-specific predictions.

The Multi-Modal Feature Fusion architecture, shown in Figure 6, presents the integration of the complementary capabilities of visual and skeletal modalities. This design leverages the broader contextual understanding provided by visual input alongside the fine-grained joint-level insights derived from pose features. Instead of relying solely on raw concatenation, the architecture adopts a staged transformation process, where each successive layer serves to refine and compress the joint feature space, selectively retaining the most informative cross-modal patterns. The classification head produces independent probability estimates for each assessment criterion, enabling detailed feedback by accounting for the fact that different aspects of movement correctness can occur independently.

3.5 Training Strategy and Implementation

A participant-based data splitting strategy was implemented to ensure robust evaluation and prevent data leakage. This approach segregated participants (rather than individual exercise repetitions) into train/validation/test sets with ratios of 70/15/15. This methodology is crucial for realistic performance as-

assessment as it prevents the model from recognizing specific individuals' movement patterns across data splits, ensuring generalization to unseen participants. Implementation utilized PyTorch with CUDA acceleration on NVIDIA P100 GPUs. Models were trained with a batch size of 16, optimized using Adam with an initial learning rate of $1e-4$ and weight decay of $1e-5$. Dynamic learning rate adjustment was implemented through ReduceLROnPlateau scheduling, reducing the learning rate by a factor of 0.5 when validation loss plateaued for 5 consecutive epochs. Binary cross-entropy with logits loss served as the primary objective, complemented by gradient clipping (max norm 1.0) and early stopping with a patience of 10 epochs to prevent overfitting.

4 RESULTS AND ANALYSIS

4.1 Evaluation Metrics

In this study, two primary metrics were utilized for evaluating model performance: Hamming Loss and F1-score. Hamming Loss quantifies classification errors in multi-label tasks by measuring the fraction of incorrectly predicted labels, where the labels represent the biomechanical criteria specific to each exercise type:

$$HL = \frac{1}{N \cdot L} \sum_{i=1}^N \sum_{j=1}^L \mathbb{I}(y_{ij} \neq \hat{y}_{ij}) \quad (3)$$

where N represents the number of samples, L denotes the number of labels, y_{ij} and $\hat{y}_{ij}$ are true and predicted labels, and $\mathbb{I}(\cdot)$ is the indicator function. Lower values indicate superior performance. F1-score was employed to assess precision-recall balance:

$$F1\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

4.2 Comparative Analysis of Model Architectures

Table 4 presents a comprehensive cross-architecture performance comparison, including both our proposed multi-modal approach and several strong baseline models implemented specifically for controlled benchmarking. The results demonstrate multiple significant findings across model architectures. First, our proposed Multi-Modal architecture consistently outperforms all alternatives across all three exercise types, achieving average Hamming Loss reductions of

1.5%, 24.2%, and 12.3% compared to the best competing approach (1D CNN-GRU + 3D ResNet (Chen et al., 2023; Du et al., 2023)).

Among single-stream approaches, our biomechanical feature engineering strategy substantially improved pose-based assessment, with combined angle and distance features reducing Hamming Loss by 32.6% for squats compared to raw landmarks. The Vision-based model achieved a 65.5% lower Hamming Loss than the optimized Pose-based model for squats (0.0370 vs. 0.1074), demonstrating the strength of 3D convolutional networks (Du et al., 2023) in capturing visual cues, which skeletal data miss. The superior performance of the Multi-Modal model reflects its ability to fuse detailed biomechanical features with broader visual context, offering a more complete view of exercise quality.

4.3 Biomechanical Criteria Analysis

The tables (5, 6, 7) analyze model performance across specific exercise criteria, revealing distinct modality capabilities.

Biomechanical criteria analysis reveals distinct, complementary strengths across modalities. In squat assessment (Table 5), the Pose-based backbone excels at precise joint relationships, achieving perfect classification for "hip/knee simultaneous bend" (HL: 0.0000), while the Vision-based backbone demonstrates superior capability for contextual understanding, perfectly classifying "hips moving backward," "neutral lower back," and "hips below knee level." This dichotomy illustrates how skeletal representations effectively encode explicit biomechanical relationships, while 3D convolutions capture holistic movement qualities that transcend landmark data.

In the deadlift assessment (Table 6), each biomechanical criterion highlights modality-specific strengths. The Pose-based model excels in "Balance" (HL: 0.1220 vs. 0.1463 for Vision), suggesting skeletal landmarks better capture stability. Conversely, "Control reverse" sees an 83% error reduction from Pose to Vision (HL: 0.2927 to 0.0488), emphasizing Vision's advantage in modeling temporal dynamics. Vision also achieves perfect classification (HL: 0.0000) for "Straight back" and "Knee bent," key for injury prevention. For "Leg extension," the Multi-Modal model improves on Vision (25% error reduction), showing the benefit of integrating both modalities for complex joint and motion assessments.

The lunge exercise (Table 7) highlights this complementarity most dramatically, with the Vision-based approach achieving 90% error reduction for "back knee position" compared to the Pose-based ap-

Table 4: Cross-architecture performance comparison. The single-stream models and multi-modal approaches shown here represent custom baseline implementations developed specifically for this study to enable controlled comparison against our proposed architecture.

Model	Squat		Deadlift		Lunges	
	HL	F1	HL	F1	HL	F1
<i>Single-Stream Models</i>						
Pose (Raw)	0.1593	0.8232	0.2000	0.7232	0.2689	0.6814
Pose (Distances)	0.1519	0.8211	0.1902	0.7054	0.2689	0.6949
Pose (Angles)	0.1296	0.8502	0.1805	0.7099	0.2521	0.6794
Pose (Angles+Distances)	0.1074	0.8795	0.1463	0.7865	0.2521	0.7139
Vision-Based (3D ResNet) (Du et al., 2023)	0.0370	0.9583	0.0585	0.9217	0.1008	0.8805
<i>Competing Multi-Modal Approaches</i>						
1D CNN-GRU + ViT (Chen et al., 2023)	0.0481	0.9502	0.1073	0.9276	0.1681	0.8837
2D CNN + ViT (Arnab et al., 2021)	0.0444	0.9665	0.0927	0.9377	0.1597	0.8862
1D CNN-GRU + 3D ResNet (Chen et al., 2023; Du et al., 2023)	0.0263	0.9694	0.0644	0.9303	0.0862	0.8947
Multi-Modal (Proposed)	0.0259	0.9706	0.0488	0.9347	0.0756	0.9097

Table 5: Squat biomechanical criteria Hamming loss.

Biomechanical Criterion	Pose	Vision	Multi
Feet flat	0.2000	0.0444	0.0667
Hip/knee bend	0.0000	0.0444	0.0000
Hips backward	0.1111	0.0000	0.0000
Neutral back	0.1111	0.0000	0.0000
Hips below knee	0.0222	0.0000	0.0000
Feet angled	0.2000	0.1333	0.0889
Overall	0.1074	0.0370	0.0259

Table 6: Deadlift biomechanical criteria Hamming loss.

Biomechanical Criterion	Pose	Vision	Multi
Balance	0.1220	0.1463	0.1220
Straight back	0.0244	0.0000	0.0000
Leg extension	0.1951	0.0976	0.0732
Control reverse	0.2927	0.0488	0.0488
Knee bent	0.0976	0.0000	0.0000
Overall	0.1463	0.0585	0.0488

proach (HL: 0.0588 vs. 0.5294)—demonstrating vision’s advantage where landmark detection proves challenging. Similarly, for “head looking forward,” the vision approach reduces error by 80% (HL: 0.0588 vs. 0.2941) by capturing orientation cues poorly represented in skeletal models.

The Multi-Modal architecture effectively leverages these complementary strengths, consistently outperforming individual modalities across biomechanical criteria. For complex criteria like “feet angled outward” in squats, it delivers 33.3% improvement over the Vision-based model and 55.6% over the Pose-based model (HL: 0.0889 vs. 0.1333 and 0.2000), while achieving perfect classification for “head looking forward” in lunges where both individual modalities exhibit errors.

Table 7: Lunge biomechanical criteria Hamming loss.

Biomechanical Criterion	Pose	Vision	Multi
Heels width	0.2941	0.2353	0.2353
Head forward	0.2941	0.0588	0.0000
90° knee bend	0.2353	0.1176	0.0588
Arms swing	0.1765	0.0588	0.0588
Feet aligned	0.0588	0.0588	0.0588
Back straight	0.1765	0.1176	0.0588
Back knee pos.	0.5294	0.0588	0.0588
Overall	0.2521	0.1008	0.0756

4.4 Computational Efficiency Analysis

Edge deployment feasibility was evaluated on a Raspberry Pi 4 (Quad-core Cortex-A72 @ 1.8GHz, 4GB RAM) to simulate real-world portable applications like mobile fitness apps and smart gym equipment. Results are shown in Table 8.

Table 8: Inference Performance Metrics.

Model	Inference Time (s)
Pose-Based	0.0192
Vision-Based	5.2748
Multi-Modal	5.6324

The Pose-based backbone demonstrates remarkable computational efficiency (0.0192s per inference), enabling real-time feedback on edge devices like smartphone. In contrast, Vision-based and Multi-Modal models exhibit significantly higher latencies (5.27s and 5.63s), making them more suitable for cloud-based processing or offline analysis. This presents a practical deployment spectrum where the Pose-based model could provide immediate feedback during exercise sessions on resource-constrained devices, while the more accurate Multi-Modal approach could be employed for detailed post-workout analysis or in scenarios where higher computational resources

are available. This accuracy-efficiency trade-off offers flexibility in designing exercise assessment systems based on specific user requirements and hardware constraints.

4.5 Key Findings

The comprehensive evaluation yields several important conclusions:

- Multi-modal integration consistently outperforms single-modality and competing hybrid approaches, with average Hamming Loss reductions of 30.0% compared to the Vision-based backbone and 79.5% compared to the Pose-based backbone.
- Feature engineering significantly enhances pose-based assessment, with combined angle and distance features reducing Hamming Loss by 32.6% for squats compared to raw landmark coordinates.
- The component backbones exhibit complementary strengths: the Vision-based component excels at capturing motion context and spatial relationships, while the Pose-based component demonstrates advantages in precise joint angle assessment.
- Computational requirements vary dramatically between components, with the Pose-based backbone offering 293× faster inference than the integrated Multi-Modal system on edge devices.

5 CONCLUSION

This study presents ALEX-GYM-1, a novel multi-camera view dataset with synchronized lateral and frontal camera perspectives and criterion-specific annotations for exercise assessment, alongside a hybrid architecture that integrates vision-based and pose-based modalities for comprehensive evaluation. Experimental results demonstrate the Multi-Modal approach's superior performance, achieving 30.0% and 79.5% Hamming Loss reductions compared to Vision-based and Pose-based models respectively. Biomechanical criteria analysis revealed complementary strengths across modalities: Vision-based components excel at contextual elements (90% error reduction for back knee positioning), while Pose-based components demonstrate precision in joint relationships (perfect classification for hip/knee simultaneous bend). The computational efficiency evaluation established a deployment spectrum from lightweight Pose-based systems (0.0192s inference) suitable for real-time applications to high-accuracy Multi-Modal

architectures (5.6324s) for detailed analysis. These findings validate the hypothesis that multi-modal integration transcends single-modality limitations, establishing a new benchmark in automated exercise assessment. Future work will focus on optimizing the architecture for reduced latency, and expanding the dataset with diverse biomechanical scenarios to further enhance assessment capabilities across fitness monitoring, rehabilitation, and personalized training applications. Future work will focus on further optimizing the architecture for reduced latency and expanding the dataset to improve generalizability and performance. While the current dataset is sufficient for validating and benchmarking different architectures across three specific exercises, deploying production-ready models will require a significantly larger and more diverse dataset. This includes incorporating a wider range of exercise types and biomechanical scenarios to ensure robust performance in real-world applications.

Regarding latency optimization, while the pose-based modality already shows strong efficiency on resource-constrained hardware, further improvements will focus on the vision-based modality. With proper architectural tuning, the vision-based branch can be optimized for lower latency, enabling the use of more accurate, real-time systems even on limited hardware. This broadens the potential for deployment on edge and mobile devices

AVAILABILITY

The proposed method and the ALEX-GYM-1 dataset are publicly available. Please refer to the dataset via the citation (Hassan et al., 2025).

ACKNOWLEDGMENTS

This work is Funded by the Science and Technology Development Fund STDF (Egypt); Project id: 51399 - "VERAS: Virtual Exercise Recognition and Assessment System".

REFERENCES

- Arnab, A. et al. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.
- Baradel, F., Wolf, C., and Mille, J. (2018). Human activity recognition with pose-driven attention to rgb. In *Proc. BMVC 2018-29th Brit. Mach. Vision Conf.*

- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields.
- Capecchi, M., Ceravolo, M. G., Ferracuti, F., Iarlori, S., Monteriù, A., Romeo, L., and Verdini, F. (2019). The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 27(7):1436–1448.
- Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 168–172.
- Chen, J. et al. (2023). Vision transformer with gru for action recognition in skeleton sequences. *Neurocomputing*.
- Clark, R. A. et al. (2021). The use of wearable technology and computer vision for movement analysis and injury prevention in sport. *J. Sports Sci.*, 39(5):533–544.
- Dosovitskiy, A. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Du, X., Li, Y., Cui, Y., Qian, R., Li, J., and Bello, I. (2023). Revisiting 3d resnets for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2899.
- Duong-Trung, N., Kotte, H., and Kravcik, M. (2023). Augmented intelligence in tutoring systems: A case study in real-time pose tracking to enhance the self-learning of fitness exercises. In *Responsive and Sustainable Educational Futures*, pages 705–710. Springer.
- Ganesh, P., Idgahi, R., Venkatesh, C., Babu, A., and Kyrarini, M. (2020). Personalized system for human gym activity recognition using an rgb camera. In *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 1–7.
- Ghorbani, S., Mahdavian, K., Thaler, A., Kording, K., Cook, D. J., Blohm, G., and Troje, N. F. (2021). Movi: A large multi-purpose human motion and video dataset. *PLOS ONE*, 16(6):e0253157.
- Hassan, A., Serour, A., Gamea, A., and Gomaa, W. (2025). Alex-gym-1: A multi-camera dataset for automated evaluation of exercise form. GitHub. Accessed: 2025-07-21.
- Kaseris, M., Kostavelis, I., and Malassiotis, S. (2024). A comprehensive survey on deep learning methods in human activity recognition. *Machine Learning and Knowledge Extraction*, 6(2):842–876.
- Keogh, J. W. and Winwood, P. W. (2017). The epidemiology of injuries across the weight-training sports. *Sports Med.*, 47(3):479–501.
- Khan, I. U., Afzal, S., and Lee, J. W. (2022). Human activity recognition via hybrid deep learning-based model. *Sensors*, 22(1):323.
- Kotte, H., Kravcik, M., and Duong-Trung, N. (2023). Real-time posture correction in gym exercises: A computer vision-based approach for performance analysis, error classification and feedback. *International Journal of Artificial Intelligence in Education*.
- Neha, D. and Manju, D. (2023). Virtual fitness trainer using artificial intelligence. *International Journal for Research in Applied Science and Engineering Technology*, 11(3):1499–1507.
- Sharshar, A., Fayez, A., Eitta, A. A., and Gomaa, W. (2022). Mm-dos: A novel dataset of workout activities. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Wijekoon, A., Wiratunga, N., and Cooper, K. (2019). MEX [dataset]. <https://doi.org/10.24432/C59K6T>. UCI Machine Learning Repository.
- Winwood, P. W., Hume, P. A., Cronin, J. B., and Keogh, J. W. L. (2014). Retrospective injury epidemiology of strongman athletes. *Journal of Strength and Conditioning Research*, 28(1):28–42.
- Wushao, W., Hefu, S., and Hanxiang, Z. (2000). An epidemiological survey and comparative study of the injuries in weightlifting. *Sports Science*, 20(4):44–46.
- Youssef, F., Parque, V., and Gomaa, W. (2023). Vcoach: A virtual coaching system based on visual streaming. *Procedia Computer Science*, 222:207–216.
- Zhang, C., Liu, L., Yao, M., Chen, W., Chen, D., and Wu, Y. (2021). Hsipu2: A new human physical fitness action dataset for recognition and 3d reconstruction evaluation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pages 481–489.