IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI-DRIVEN FITNESS EVALUATION SOFTWARE FOR FITNESS ASSESSMENTS

Srinidhi B, Sowmiya R, Shridevi SR

Department of Computer Science and Engineering
Sri Venkateswara College of Engineering

DOI: https://doi.org/10.56975/ijcrt.v14i7.309012

Abstract: This paper proposes a novel approach for real-time fitness evaluation by employing computer vision and deep learning technologies. Existing fitness monitoring systems typically rely on human supervision or wearable sensors, which are often invasive, expensive, and inaccessible for many users. The proposed system uses pose estimation to extract skeletal keypoints from video input and processes them as temporal sequences using Bidirectional Long Short-Term Memory (BiLSTM) models. An automatic labeling strategy based on joint angle thresholds is employed to classify exercise stages, and a multi-model strategy is used to independently handle push-ups, squats, and pull-ups. The system combines classification and detection approaches to accurately identify exercise types and count repetitions by detecting transitions between motion stages. Experimental results demonstrate high accuracy in exercise classification, stage detection, and repetition counting, confirming the potential of the proposed approach as a non-intrusive and efficient solution for real-time fitness evaluation.

Index Terms: pose estimation, BiLSTM, fitness monitoring, exercise recognition, computer vision, deep learning

I. INTRODUCTION

The growing interest in health and fitness has driven demand for intelligent systems capable of assessing and measuring physical exercise performance with precision. Traditional fitness evaluation methods typically require manual supervision by trained professionals or the use of wearable sensors, both of which introduce significant barriers — including cost, intrusiveness, and limited scalability across diverse user populations. The widespread availability of high-resolution cameras combined with advances in computer vision and deep learning now makes it possible to perform non-intrusive, sensor-free fitness monitoring from standard video footage.

This paper presents an Artificial Intelligence-based fitness assessment system that integrates pose estimation with deep learning sequence models to automatically identify exercise types, detect motion stages, and track repetition counts in real-time. The system employs the MediaPipe Pose Landmarker to extract 33 skeletal landmarks from each video frame, producing a 66-dimensional spatial feature vector representing body configuration. These per-frame feature vectors are assembled into temporal sequences and fed into a trained Bidirectional Long Short-Term Memory (BiLSTM) network, which captures motion dynamics in both forward and backward temporal directions.

Unlike prior approaches that rely solely on classification to determine exercise type, the proposed system employs both a classification stage (to identify the exercise category) and a detection stage (to count repetitions based on stage transitions). The system's multi-model design — where separate BiLSTM models are trained for each exercise — enables each model to specialize in the unique motion patterns of its respective exercise, improving robustness and reducing inter-class confusion. Additionally, the system uses

automatically generated labels derived from joint angle thresholds, reducing the need for manual annotation and improving the scalability of the training pipeline.

II. RELATED WORK

A. Deep Learning for Human Activity Recognition

Research in human activity recognition (HAR) using deep learning has progressed significantly over the past decade, moving from sensor-based approaches to video and pose-based methods. Early work by Shi et al. [1] demonstrated the effectiveness of Convolutional LSTM (ConvLSTM) models for learning spatiotemporal dependencies in sequential data. The model simultaneously captured spatial structure via convolutional filters and temporal dependencies through LSTM recurrence, establishing a strong foundation for video-based activity recognition.

Ordóñez and Roggen [2] proposed a hybrid architecture combining convolutional and LSTM layers for multimodal wearable activity recognition, showing that jointly modeling spatial features and temporal patterns from multi-sensor streams yields superior performance over architectures that treat each modality independently. Hammerla et al. [3] conducted a comprehensive benchmark comparing convolutional, recurrent, and hybrid deep learning models on activity recognition benchmarks, concluding that recurrent architectures are particularly effective when motion data exhibits strong temporal structure.

B. Temporal Modeling for Activity Recognition

Temporal modeling is fundamental to understanding complex, continuous human motion. Wang et al. [4] provided a thorough review of sensor-based activity recognition techniques using deep learning, emphasizing that the capacity to model long-range temporal dependencies is critical for differentiating similar activity patterns. Standard unidirectional LSTM models process sequences in a single temporal direction, limiting their ability to leverage future context — a significant drawback in exercise analysis where both the preceding and following frames inform the current stage label. Bidirectional LSTMs address this limitation by simultaneously learning from past and future temporal context, improving recognition accuracy for activities with symmetric motion profiles such as push-ups and squats.

C. Pose-Based Fitness Analysis Using Computer Vision

Pose estimation has emerged as a key enabler for vision-based fitness monitoring. MediaPipe, introduced by Lugaresi et al. [5], provides a lightweight, real-time framework for extracting skeletal keypoints from video, enabling non-intrusive monitoring without specialized hardware. However, the majority of pose-based fitness systems rely on frame-level thresholding or simple rule-based classification, which fails to capture the rich temporal dynamics of sustained exercise. These approaches are particularly vulnerable to noise in transition frames and are unable to generalize across users with different body proportions or motion speeds. The proposed system addresses these limitations by coupling MediaPipe's pose estimation with BiLSTM-based sequence modeling, enabling robust temporal reasoning over sliding windows of pose sequences.

III. METHODOLOGY

A. System Overview

The proposed AI-based fitness assessment system is structured around three interconnected stages: (1) pose estimation and feature extraction from video frames, (2) temporal sequence-based model training using BiLSTM networks, and (3) real-time inference for exercise identification, stage classification, and repetition counting. The system is designed to operate purely on video input, requiring no additional hardware or specialized sensors.

B. Pose Estimation and Feature Extraction

The MediaPipe Pose Landmarker model is used to detect 33 skeletal keypoints per frame, covering major anatomical joints including shoulders, elbows, wrists, hips, knees, and ankles. Each frame is converted from BGR to RGB color space before processing to ensure compatibility with the pose estimation model. The normalized (x, y) coordinates of all 33 landmarks are concatenated to produce a 66-dimensional feature vector representing the spatial configuration of the body at a given instant. This representation eliminates pixel-level image dependency, reducing computational overhead and enabling the model to generalize across different camera distances and body sizes. Joint angles are computed from three-point geometric relationships between adjacent landmarks (e.g., shoulder–elbow–wrist for elbow angle), providing biomechanically meaningful features that complement the raw coordinate representation.

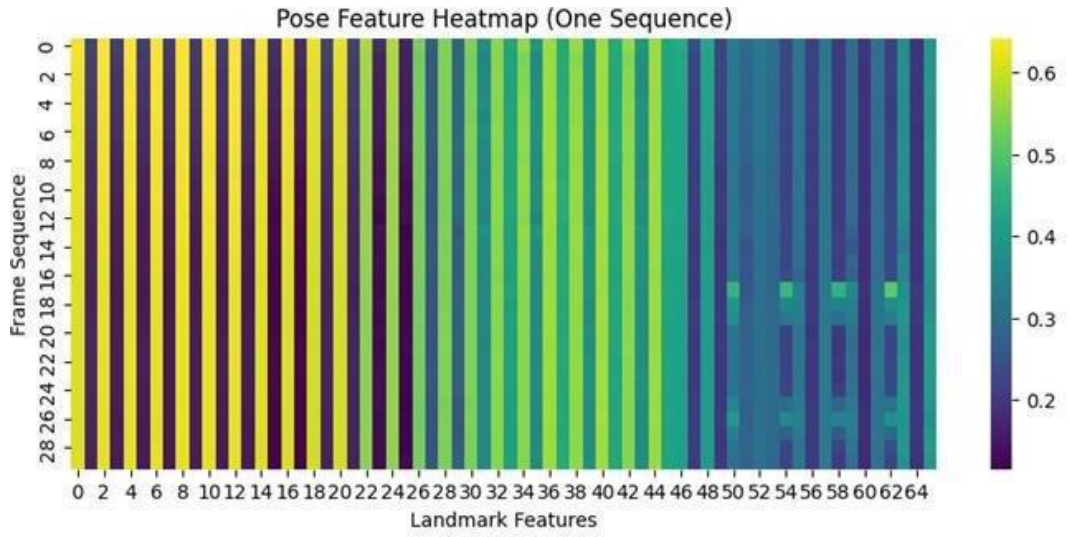

fig. 1. pose feature heatmap over a single temporal sequence (30 frames × 66 landmark features).

C. Automatic Data Labeling

Manual frame-level annotation of exercise videos is both time-consuming and prone to inconsistency. To address this, the proposed system employs a rule-based automatic labeling scheme grounded in joint angle thresholds. Each frame is assigned one of two binary labels — UP (1) or DOWN (0) — based on the computed joint angle for the relevant exercise. For squats, a knee angle below 100° is labeled DOWN and above 160° is labeled UP. For push-ups, an elbow angle below 90° is labeled DOWN and above 150° is labeled UP. For pull-ups, labeling incorporates both arm angle and chin position: an angle above 160° with the chin below shoulder level is labeled DOWN, while an angle below 90° with the chin above shoulder level is labeled UP. Frames with ambiguous intermediate angles are discarded to prevent noise in the training set and improve the purity of the learned motion representations.

D. Temporal Sequence Generation

Exercise recognition is inherently a temporal task — the meaning of a given pose is context-dependent and cannot be inferred from a single frame in isolation. To model temporal dynamics, the proposed system applies a sliding window approach over the sequence of per-frame feature vectors. Each window spans 30 consecutive frames and is assigned the label of its final frame, effectively encoding the motion trajectory leading up to a particular stage. This produces a time-series dataset suitable for sequence-to-label learning. The 30-frame window length was empirically selected to balance temporal coverage with computational efficiency during real-time inference.

E. Model Architecture

The proposed BiLSTM model accepts a 30-frame sequence of 66-dimensional feature vectors as input (shape: 30 × 66). The network consists of two BiLSTM layers with 64 and 32 units respectively, enabling the model to learn hierarchical temporal representations while progressively compressing the sequence into a compact feature descriptor. A fully connected dense layer with ReLU activation follows, introducing non-linearity and enabling the model to learn complex stage-discriminative patterns. A dropout rate of 0.3 is applied after the dense layer to regularize the model and reduce overfitting. The output layer is a softmax classifier that produces a probability distribution over two classes: UP (1) and DOWN (0). The complete system architecture, including the pose estimation pipeline, BiLSTM network, decision logic, and output layer, is illustrated in Fig. 2.

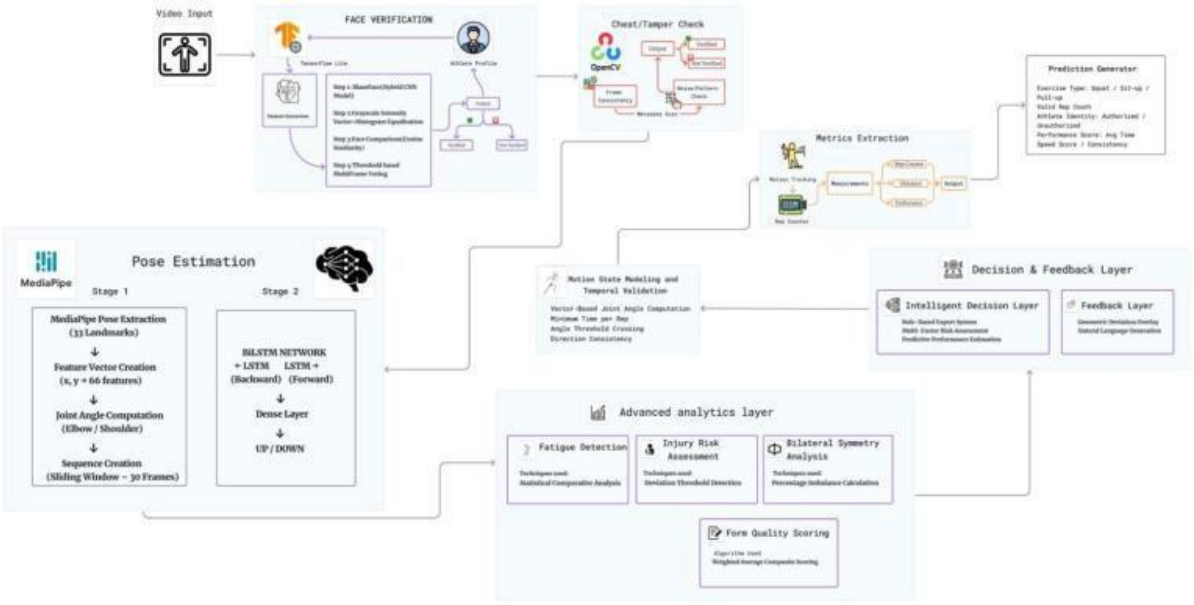

fig. 2. complete system architecture: pose estimation, bilstm sequence modeling, decision and feedback layers.

F. Multi-Model Training Strategy

Rather than training a single generalized model across all exercise types, the proposed system employs a multi-model strategy in which a dedicated BiLSTM model is trained independently for each exercise category (push-ups, squats, and pull-ups). Although all three models share the same architectural design, they are trained on exercise-specific datasets, allowing each model to specialize in the distinctive biomechanical motion patterns of its respective exercise. This approach significantly reduces inter-class confusion and improves stage classification accuracy. All models are trained using the Adam optimizer with sparse categorical cross-entropy loss, with a 20% validation split for performance monitoring. Automatic labeling and sliding window sequence generation are applied uniformly across all training datasets.

G. Real-Time Inference and Exercise Detection

During deployment, the system processes each incoming video frame in real-time. Skeletal landmarks are extracted and assembled into a rolling 30-frame sequence buffer. Once the buffer is populated, the feature sequence is passed to all three trained exercise models simultaneously. Each model outputs a probability distribution over the UP and DOWN stages, from which a confidence score is derived. The exercise with the highest confidence score is selected as the active exercise type according to:

$$Exercise = argmax(C_{pushup}, C_{squat}, C_{pullup})$$

where C represents the confidence score for each model. This confidence-based selection mechanism leverages model specialization to achieve robust exercise identification under varying conditions. The sliding window buffer additionally provides temporal smoothing, reducing prediction instability at stage transition boundaries.

fig. 3. real-time exercise detection output overlaid on video frame, showing exercise type, current stage, and repetition counters.

H. Repetition Counting Mechanism

Repetition counting is implemented by monitoring stage transitions in the model's sequential predictions. Each exercise is characterized by two alternating stages (UP and DOWN), and a complete repetition requires traversal of the full UP → DOWN → UP (or DOWN → UP → DOWN) cycle. For push-ups and squats, a repetition is counted upon the transition from DOWN to UP, reflecting the completion of the exertion phase. For pull-ups, a repetition is counted upon the transition from UP to DOWN, reflecting the controlled descent. The mechanism maintains a record of the previous stage and increments the counter only upon a valid complete-cycle transition, preventing duplicate counts from sustained or ambiguous poses.

fig. 4. repetition counting mechanism based on stage transition detection between up and down positions.

IV. EXPERIMENTAL RESULTS

A. Classification Performance

The performance of each BiLSTM model was evaluated on held-out test sequences for the respective exercise. All three models demonstrated high accuracy in classifying the UP and DOWN motion stages, reflecting the effectiveness of the temporal sequence representation in capturing biomechanical patterns. The push-up model achieved an overall classification accuracy of 95%, with precision of 0.92 and recall of 0.81 for the DOWN class, and precision of 0.96 and recall of 0.99 for the UP class. The weighted average F1-score across both classes was 0.95, confirming consistently strong performance across the full test set.

B. Confusion Matrix Analysis

The confusion matrix for the push-up model (Fig. 5) illustrates that the vast majority of test samples are correctly classified along the diagonal. Of 1,192 actual DOWN samples, 39 were incorrectly predicted as UP. Of 3,753 actual UP samples, 57 were incorrectly classified as DOWN. The misclassified samples predominantly correspond to transitional poses at stage boundaries where joint angles fall near the labeling threshold, representing inherently ambiguous frames rather than systematic model errors. The low off-diagonal counts confirm that the BiLSTM model generalizes effectively across different users and movement speeds.

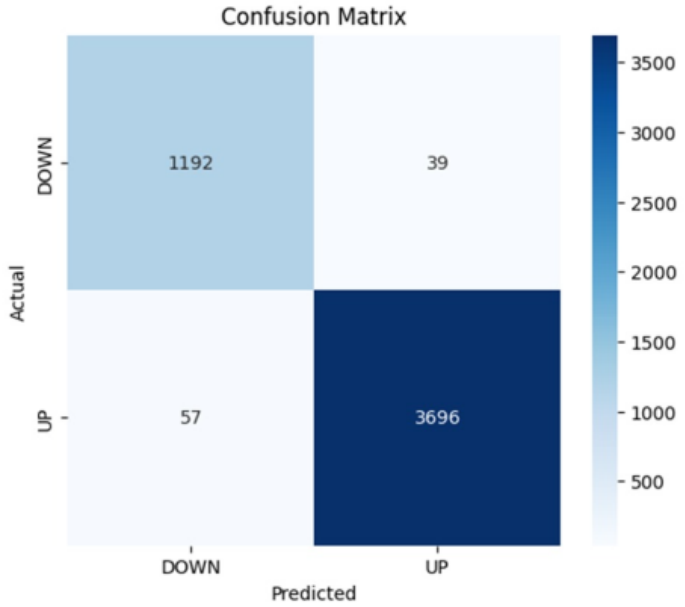

fig. 5. confusion matrix for the push-up bilstm classification model.

C. ROC Curve and AUC Metrics

The Receiver Operating Characteristic (ROC) curve for the squat classification model is shown in Fig. 6. The model achieves an AUC of 0.99, indicating near-perfect discrimination between UP and DOWN stages across all decision thresholds. This result demonstrates that the model maintains a strong balance between sensitivity (true positive rate) and specificity (1 − false positive rate), confirming robust generalization beyond the training distribution. The high AUC is consistent with the model's ability to leverage joint angle features that provide clear, biomechanically grounded separation between the two motion stages.

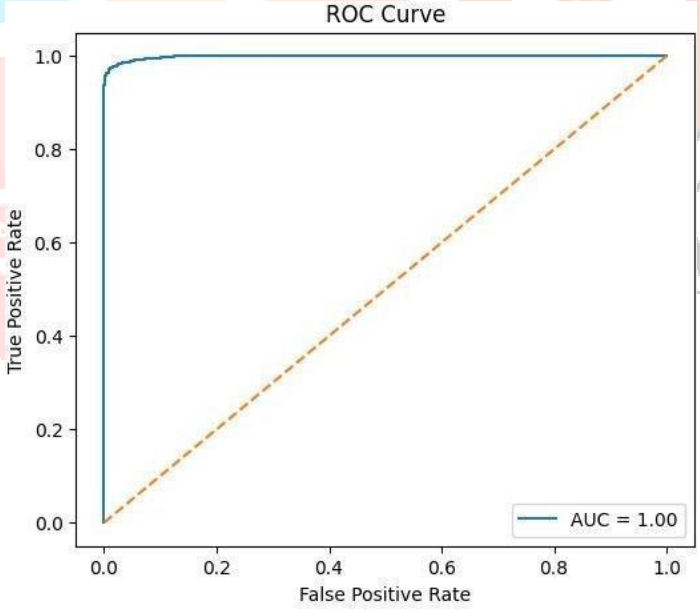

fig. 6. roc curve for the squat classification model (auc = 0.99).

D. Feature Importance Analysis

To understand which input features most strongly drive classification decisions, a logistic regression feature importance analysis was conducted on the push-up model's learned representations. As illustrated in Fig. 7, features corresponding to elbow and wrist joint coordinates — particularly the y-axis positions of the right and left elbows — exhibit the largest absolute coefficient values, reflecting their strong association with the UP/DOWN stage distinction during push-up execution. Features associated with the lower body (hip, knee) show comparatively smaller contributions, consistent with the expectation that push-up stage discrimination is primarily driven by upper-limb kinematics. These findings validate the system's reliance on joint angle features and confirm that incorporating both raw keypoint coordinates and derived angular features enhances model interpretability.

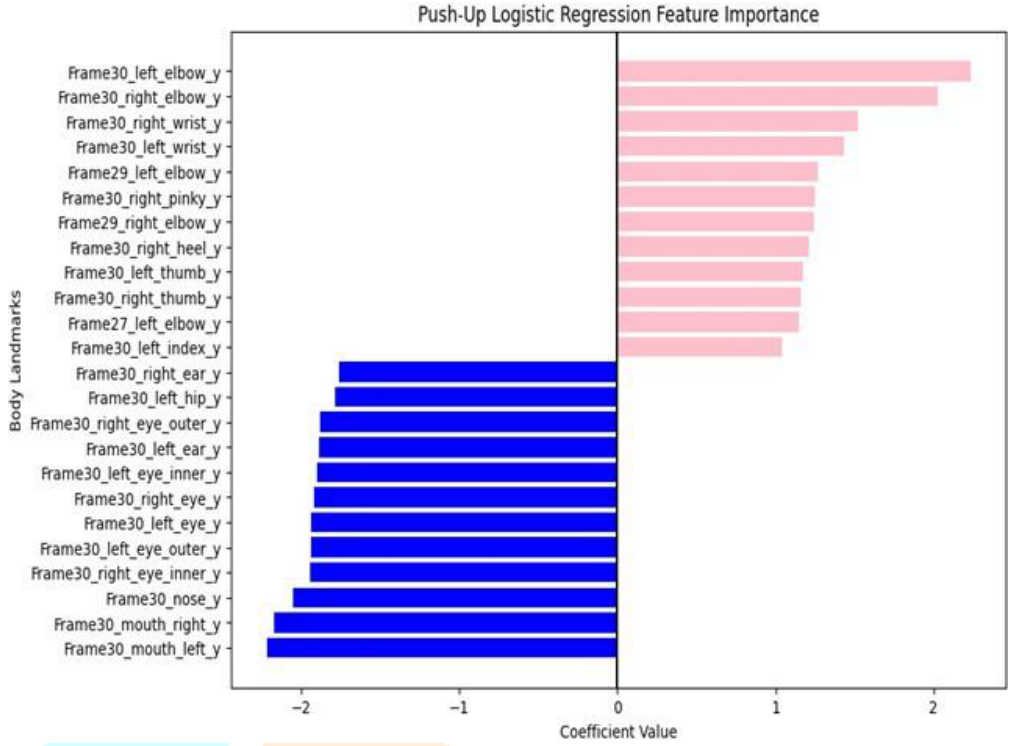

fig. 7. feature importance analysis showing coefficient magnitudes for up and down stage classification.

E. Classification Report

Classification Report:

	precision	recall	f1-score	support
DOWN	0.92	0.81	0.86	105
UP	0.96	0.99	0.97	486
accuracy			0.95	591
macro avg	0.94	0.90	0.92	591
weighted avg	0.95	0.95	0.95	591

fig. 8. per-class precision, recall, f1-score, and support for the push-up model (accuracy = 0.95, n = 591).

F. Repetition Counting Accuracy

The repetition counting mechanism was validated across multiple trial videos for each exercise type. The system consistently produced accurate repetition counts without duplicate increments or missed transitions across all evaluated sequences. By requiring full UP → DOWN → UP cycle completion before incrementing the counter, the mechanism is inherently robust to brief pauses, slow execution, or partial repetitions. This design ensures that the system counts only biomechanically meaningful complete repetitions, which is critical for fitness assessment applications where accurate load quantification is required.

G. Real-Time System Performance

The system was deployed and evaluated on live and recorded video streams. Frame-level pose estimation and feature extraction operated efficiently within real-time constraints, with the sliding window buffer providing smooth, temporally consistent predictions. No additional hardware beyond a standard camera was required, confirming the system's suitability for accessible, low-cost fitness monitoring applications.

V. DISCUSSION

The experimental results confirm that the proposed system effectively addresses the limitations of existing fitness monitoring approaches by providing accurate, non-intrusive exercise evaluation from standard video input. The integration of MediaPipe pose estimation with BiLSTM sequence modeling enables the system to capture both the spatial configuration of the body and the temporal dynamics of motion — two dimensions that are jointly necessary for robust exercise recognition. The multi-model strategy further improves performance by allowing each model to focus on the specific motion patterns of a single exercise type, reducing the risk of cross-exercise confusion.

Compared to wearable sensor-based systems, the proposed approach eliminates the requirement for body-worn devices, reducing user burden and enabling deployment in any environment with a camera. Compared to frame-level threshold methods, the BiLSTM-based approach is more robust to motion variability, body proportion differences, and execution speed variations. The automatic labeling system significantly reduces the cost of dataset creation, making the pipeline scalable to new exercise types.

Despite these strengths, several limitations warrant consideration. The system's performance is sensitive to camera placement and angle — significant occlusion or extreme viewpoints may degrade pose estimation quality and reduce classification accuracy. Illumination conditions also affect the reliability of landmark detection. The current system supports three exercise types; extending coverage to a broader exercise vocabulary would require additional training data and potentially exercise-specific labeling rules. Future improvements may include adaptive thresholding for labeling, multi-view input fusion, and user-specific calibration to improve generalization across body types and fitness levels.

VI. CONCLUSION

This paper has presented an AI-driven fitness assessment system that combines MediaPipe-based pose estimation with Bidirectional Long Short-Term Memory sequence modeling to enable accurate, real-time exercise recognition, stage classification, and repetition counting from video input alone. The system's multi-model architecture — with dedicated BiLSTM models for push-ups, squats, and pull-ups — and its automatic joint-angle-based labeling pipeline collectively enable high classification accuracy (up to 95%) and near-perfect stage discrimination (AUC = 0.99) without requiring wearable sensors or manual annotation. The proposed system represents a scalable, accessible, and non-intrusive foundation for intelligent fitness monitoring, with clear pathways for extension to broader exercise vocabularies, mobile deployment, and real-time posture correction feedback.

VII. FUTURE WORK

Several directions for future development are identified. First, the system could be extended to support a wider variety of exercises — including lunges, deadlifts, and overhead presses — by collecting additional labeled video data and training corresponding exercise-specific models. Second, real-time posture correction feedback could be integrated by defining biomechanical quality thresholds beyond simple UP/DOWN stage classification, enabling the system to detect form errors such as knee valgus during squats or elbow flare during push-ups. Third, robustness under challenging conditions — including low illumination, partial occlusion, and non-frontal camera angles — could be improved through data augmentation and multi-view pose fusion. Finally, deployment on mobile devices via model compression techniques such as knowledge distillation or quantization would make the system practical for personal fitness monitoring in everyday environments.

REFERENCES

[1] X. Shi et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," Advances in Neural Information Processing Systems (NeurIPS), 2015.

[2] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," Sensors, vol. 16, no. 1, p. 115, 2016.

[3] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables," in Proc. IJCAI, 2016.

[4] J. Wang et al., "Deep Learning for Sensor-based Activity Recognition: A Survey," Pattern Recognition Letters, vol. 119, pp. 3–11, 2019.

[5] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.