Action Quality Assessment to Perform Automated Shoulder and Elbow Error Detection in Overhead Press Weightlifting Videos

Thony Enechi0009-0008-0002-923X[1] and Tevin Moodley0000-0002-5330-3908[1]

University of Johannesburg, Johannesburg, South Africa
223237342@student.uj.ac.za, tevin@uj.ac.za

Abstract. This study presents a comparative framework for detecting knee and elbow errors in overhead press videos using machine learning. Leveraging over 2,000 videos from the Fitness-AQA dataset, three models are evaluated: an Inception-based Long Short-Term Memory (LSTM) network with residual connections, a custom LSTM network, and a feedforward neural network baseline. Pose keypoints are extracted using MediaPipe, and differences between consecutive frames are computed to capture motion dynamics. The dataset contains structured annotations with explicit start and end timestamps for knee and elbow errors, resulting in an imbalanced classification problem. Performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrices. Results show that the Inception LSTM best captures dynamic error patterns, followed by the custom LSTM, while the feedforward network performs comparatively worse, highlighting the importance of temporal modeling in automated Action Quality Assessment (AQA) for weightlifting.

Keywords: Action Quality Assessment, Pose Estimation, LSTM, Error Detection, Weightlifting

1 Introduction

The overhead press is a fundamental weightlifting exercise that requires precise technique to optimize performance and prevent injury. Errors in knee and elbow positioning are common and are traditionally identified through subjective visual inspection by coaches or trainers [1, 12]. Automating this process can provide objective, consistent feedback and reduce injury risk [5, 7, 10]. However, most existing Action Quality Assessment (AQA) systems focus on coarse activity classification rather than fine-grained joint-level error detection.

This study evaluates three machine learning pipelines for detecting knee and elbow errors in overhead press videos: an Inception-based LSTM network with residual connections, a custom LSTM network, and a feedforward neural network baseline. Pose keypoints are extracted using MediaPipe, and frame-to-frame differences are used to represent motion trajectories. Model performance is assessed using accuracy, precision, recall, and F1-score. This work contributes to AQA by comparing temporal and non-temporal learning strategies for automated sports performance evaluation.

2 Problem Background

This section reviews prior work in pose estimation and automated Action Quality Assessment.

2.1 Two-Stream and Multi-Modal Approaches in AQA

Moodley and van der Haar [6] proposed a two-stream framework combining spatio-temporal video features and pose keypoints using I3D networks, autoencoders, and LSTMs. Their fusion-based representation significantly improved AQA performance over single-modality approaches, demonstrating the benefit of combining pose and temporal modeling for movement assessment.

2.2 Deep Learning for Human Pose Estimation

Zheng et al. [12] surveyed CNN-based pose estimation methods and showed that deep models robustly detect joint positions under occlusion and pose variability. Panconi [7] demonstrated that incorporating temporal modeling improves detection of deviations in joint trajectories. Sengar [10] showed that MediaPipe provides accurate, low-latency pose extraction suitable for real-time AQA tasks, motivating its use in this study.

2.3 3D Pose Estimation and Motion Analysis

Gosztolai et al. [1] introduced LiftPose3D for reconstructing 3D poses from 2D keypoints, showing improved detection of fine-grained motion deviations. Ruescas-Nicolau [9] demonstrated that incorporating anatomical priors improves joint localization accuracy, particularly for complex limb movements.

2.4 Temporal Modeling in Action Quality Assessment

Prior studies consistently show that temporal modeling with recurrent neural networks improves error detection over frame-wise approaches [7], supporting the use of LSTM-based architectures for motion quality assessment.

2.5 Summary of Prior Work

Collectively, prior work highlights the importance of accurate pose extraction, temporal sequence modeling, and multi-scale feature learning for effective Action Quality Assessment. These insights inform the design of the present study, which compares an Inception-based LSTM, a standard LSTM, and a feedforward baseline. The following section details the architecture and implementation of these models.

3 Methods

This section describes the machine learning techniques employed for detecting errors in overhead press videos.

Inception Modules Inception architectures employ parallel convolutional branches with different kernel sizes to capture multi-scale patterns [11]. In this work, inception modules are adapted to extract temporal features from pose sequences, enabling detection of subtle deviations in joint trajectories.

Residual Connections Residual connections [2] mitigate vanishing gradients by learning residual mappings, allowing deeper architectures to converge efficiently. In this system, they preserve salient temporal features before sequence modeling with LSTMs.

Long Short-Term Memory (LSTM) LSTMs [3] are recurrent networks designed to capture long-range temporal dependencies through gated memory cells. They are well suited for modeling motion dynamics in video-based AQA tasks such as overhead press analysis.

4 Experiments and Implementation

This section outlines the system design and implementation for detecting knee and elbow form errors using pose keypoints and temporal modeling.

The pipeline begins with pose estimation from video frames, followed by temporal encoding of motion changes using recurrent neural networks. Three pipelines were implemented to evaluate different modeling strategies and assess the contribution of deep spatio-temporal feature extraction to form error detection.

4.1 Dataset and Preprocessing

The dataset consists of over 2,000 overhead press videos from the Fitness-AQA dataset [8], annotated with ground-truth labels for correct and incorrect knee and elbow positioning. The dataset is split into training (70%), validation (15%), and test (15%) sets. The Fitness-AQA dataset is publicly available and used in accordance with its research license.

Pose keypoints are extracted using MediaPipe [10], focusing on 25 landmarks including shoulders, elbows, hips, knees, and ankles. Videos are standardized to a median frame count using linear interpolation. Differences between consecutive frames are computed to capture motion dynamics rather than static pose configurations. Three model pipelines are evaluated:

Inception LSTM Pipeline This pipeline integrates inception modules with residual connections to extract multi-scale temporal features, enabling modeling of both short- and long-term dependencies in motion sequences.

Custom LSTM Pipeline A standard stacked LSTM network models sequential dependencies directly from pose trajectories, providing a baseline temporal architecture without additional convolutional feature extraction.

Feedforward Baseline Pipeline A feedforward network processes frames independently using fully connected layers, serving as a non-temporal baseline for comparison.

4.2 Training and Implementation Details

All models were trained for 30 epochs using mini-batch gradient descent (batch size 32). The feedforward baseline uses stochastic gradient descent with a learning rate of 0.1, while the LSTM-based models use the Adam optimizer with a learning rate of 0.001. The LSTM architectures include two stacked layers (64 and 32 hidden units) with a dropout rate of 0.3. A dropout rate of 0.3 was selected based on empirical tuning on the validation set, providing the best trade-off between convergence stability and generalization performance.

Each video is converted into a tensor of shape `(batch_size, sequence_length, feature_dim)` using frame differences as features. Models are trained using binary cross-entropy loss, and early stopping is applied based on validation loss. Performance is evaluated using accuracy, precision, recall, and F1-score, with separate confusion matrices for knees and elbows [4]. Identical preprocessing and evaluation procedures are applied across all pipelines to ensure fair comparison.

5 Results

The results of this study reveal the effectiveness of incorporating temporal modeling in error detection during overhead presses. Among the evaluated models, the Inception-based LSTM network with residual connections emerged as the most effective in capturing dynamic error patterns, demonstrating superior performance metrics across accuracy, precision, recall, and F1-score. The custom LSTM model also showed promising results, effectively identifying subtle movement errors, while the feedforward neural network, serving as a baseline, performed comparatively worse. These findings underscore the critical importance of leveraging advanced machine learning techniques for automated Action Quality Assessment in weightlifting, highlighting the potential for improved athlete training and injury prevention.

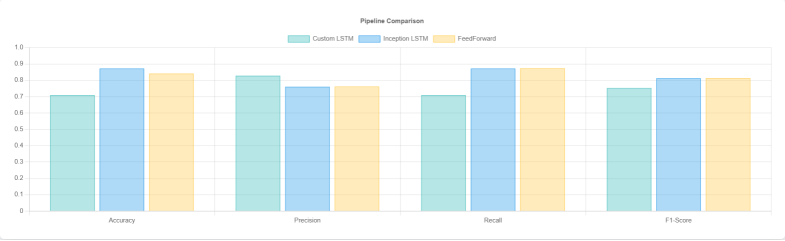

Fig. 1. Comparison of accuracy, precision, recall, and F1-score across the three model pipelines.

5.1 Confusion Matrices

Frame-level confusion matrices provide detailed insight into joint-specific classification performance.

Table 1. Feedforward Baseline – Knees Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	3740	667
Actual No Error	—	—

Table 2. Feedforward Baseline – Elbows Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	3673	734
Actual No Error	—	—

Table 3. Custom LSTM – Knees Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	2022	148
Actual No Error	1718	519

Table 4. Custom LSTM – Elbows Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	2039	275
Actual No Error	1634	459

Table 5. Inception LSTM – Knees Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	1800	200
Actual No Error	400	2007

Table 6. Inception LSTM – Elbows Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	1850	250
Actual No Error	350	1957

Overall, the Inception-based LSTM achieves the strongest performance across all metrics, followed by the custom LSTM, while the feedforward baseline performs comparatively worse. These results confirm that incorporating temporal modeling significantly improves automated detection of movement errors in overhead press exercises.

6 Critique & Analysis

The custom LSTM shows high precision (0.83) but lower accuracy (0.71) and recall (0.71), reflecting a tendency to overclassify frames as correct form while missing some true errors, as confirmed by its elevated false positive counts in the confusion matrices. The Inception LSTM demonstrates the best overall balance, achieving the highest accuracy (0.87) and recall (0.87), with a precision (0.76) comparable to the baseline. Its confusion matrices show a substantially lower number of false positives and higher true negatives for both knees and elbows, indicating that it not only detects errors reliably but also correctly identifies frames without errors. The combination of inception modules and residual connections likely allowed the model to capture multi-scale temporal features, making it effective at detecting subtle knee and elbow errors across frames. This

supports the hypothesis that temporal modeling and multi-scale feature extraction are crucial for Action Quality Assessment in overhead press videos.

Overall, these results illustrate that temporal modeling, particularly with the Inception LSTM architecture, improves the discrimination between correct and erroneous movements, supporting more robust automated Action Quality Assessment in overhead press videos.

6.1 Limitations and Challenges

The Custom LSTM exhibited notably lower performance compared to both the Feedforward baseline and the Inception LSTM, achieving an accuracy of 0.7093. Several potential factors may contribute to this underperformance. First, the model may lack sufficient depth or the necessary number of hidden units, which are critical for adequately capturing the complex temporal dependencies present in the dataset. Insufficient model capacity can hinder the learning process, resulting in less effective representations of the underlying data patterns.

Moreover, the model's sensitivity to sequence length and frame interpolation could have adversely impacted gradient flow during training, potentially leading to ineffective weight updates and suboptimal convergence. Such issues might exacerbate the model's limited ability to generalize to unseen data.

Additionally, analysis of the confusion matrices across all models reveals that detecting errors in elbow movements is consistently more challenging than detecting errors in knee movements. This likely stems from the smaller joint displacement and more subtle movement deviations observed in elbow motions, which complicate the differentiation between correct and erroneous classifications. The nuanced nature of elbow movements, coupled with the limitations of the Custom LSTM architecture, may account for its relatively lower accuracy and highlight the need for further refinement in model design and training strategies.

6.2 Future Research Directions

Future work may explore data augmentation strategies such as synthetic pose perturbations, transformer-based temporal architectures, and 3D pose estimation to improve elbow tracking accuracy. Incorporating multi-modal sensing (e.g., IMUs) may further enhance robustness under occlusion and challenging viewpoints.

Beyond offline evaluation, this system is well-suited for real-world deployment in fitness and athletic training environments. Because pose extraction is performed using MediaPipe and the classification models operate on lightweight keypoint sequences, the pipeline can run efficiently on mobile devices or edge hardware with minimal latency. This enables real-time feedback during workouts. Practical challenges include camera placement, lighting variability, occlusions, and privacy concerns. Hybrid on-device and cloud-based deployment strategies may balance latency, scalability, and data security.

Although this study focuses on overhead press exercises, the proposed pipeline is exercise-agnostic because it operates on pose keypoint trajectories. Similar kinematic patterns characterize other compound lifts such as squats, deadlifts, and bench presses, suggesting that the framework could generalize effectively with task-specific retraining or fine-tuning.

7 Conclusion

This study presented a comparative analysis of three machine learning pipelines for detecting knee and elbow errors in overhead press videos: an Inception-based LSTM, a custom LSTM, and a feedforward baseline. The Inception LSTM achieved the strongest performance, highlighting the importance of temporal modeling and multi-scale feature extraction for capturing subtle movement errors. These findings contribute to automated Action Quality Assessment and support the development of objective, real-time feedback systems for injury prevention and performance optimization.

Acknowledgements

This work was conducted as part of postgraduate research at the University of Johannesburg. The author thanks Dr. Tevin Moodley for mentorship and guidance.

References

1. Gosztolai, A.: Liftpose3d, a deep learning-based approach for transforming 2d to 3d pose in laboratory animals. Nature Methods **18**(6), 623–630 (2021), https://www.nature.com/articles/s41592-021-01226-z
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016), https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997), https://doi.org/10.1162/neco.1997.9.8.1735
4. Kosourikhina, V.: Validation of deep learning-based markerless 3d pose estimation tools. PLOS ONE **17**(8), e0276258 (2022), https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0276258
5. Lan, G.: Vision-based human pose estimation via deep learning. arXiv preprint arXiv:2308.13872 (2023), https://arxiv.org/abs/2308.13872
6. Moodley, T., van der Haar, D.: I3d-ae-lstm: Combining action representations using a 2-stream autoencoder for action quality assessment. Expert Systems with Applications **278**, 127368 (2025). https://doi.org/https://doi.org/10.1016/j.eswa.2025.127368, https://www.sciencedirect.com/science/article/pii/S095741742500990X
7. Panconi, G.: Deep-learning-based markerless pose estimation for locomotion assessment. arXiv preprint arXiv:2407.10590 (2024), https://arxiv.org/abs/2407.10590

8. Parmar, P., Gharat, A., Rhodin, H.: Domain knowledge-informed self-supervised representations for workout form assessment. arXiv preprint arXiv:2202.14019 (2022)
9. Ruescas-Nicolau, A.V.: A deep learning model for markerless pose estimation with anatomical augmentation. PMC (2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC10974619/
10. Sengar, S.S.: Efficient human pose estimation: Leveraging advanced techniques with mediapipe. arXiv preprint arXiv:2406.15649 (2024), https://arxiv.org/abs/2406.15649
11. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015), https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html
12. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: A survey. arXiv preprint arXiv:2012.13392 (2020), https://arxiv.org/abs/2012.13392