
Towards Safer Strength Training: Predicting Rep Failure Using Pose-Based Deep Learning

Caterina Mammola, Tomas Maillo

Abstract

Strength training to the point of muscular failure offers considerable benefits for hypertrophy and strength, but also raises injury risks and recovery concerns. We propose a deep learning solution that predicts when an individual is at risk of failing their next bicep curl repetition, enabling safer and more informed training decisions. Our approach addresses a gap in current exercise-analysis models, which typically focus on rep counting or form correction without considering the moment of failure. We collect a novel dataset of bicep curl sessions performed to failure and feed extracted joint coordinates from individual reps into an LSTM-based model. Inspired by the Remaining Useful Life literature, our model predicts how many repetitions remain before muscular failure occurs.

1. Introduction

Strength training to failure refers to performing an exercise until the target muscles are unable to complete another movement with proper form. While this practice is common in resistance training to maximise muscle growth, it carries inherent safety risks. If muscular failure occurs unexpectedly, trainees may lose control of the weight. In exercises such as the bench press—where the load is positioned directly above the chest—this loss of control can result in severe injury or even death (Lombardi & Troxel, 2003).

This safety concern motivates the central research question of this project: Is it possible to build a system capable of predicting when a trainee is about to reach muscular failure? We hypothesise that joint motion patterns—such as changes in arm speed or range of motion—can serve as reliable predictors of impending failure. While there is substantial literature on pose estimation for exercise-related tasks like form correction (Matsui et al., 2019) and repetition counting (Dwibedi et al., 2020), to our knowledge, it has not been applied to the problem of muscular failure prediction.

To this end, we collect a dataset of 254 bicep curl sets performed to failure, segment them into individual repetitions, extract kinematic markers from joint coordinates, and train a Hierarchical Long Short-Term Memory (LSTM) model to forecast when failure is likely to occur. We frame this as a regression problem inspired by Remaining Useful

Life (RUL) estimation, with the objective of minimizing the difference between the predicted and actual number of repetitions remaining.

We chose bicep curls as the test exercise due to their practicality and safety for novice participants. However, we believe that demonstrating the feasibility of joint-tracking-based failure prediction in this low-risk context can serve as a proof of concept for extending the approach to higher-risk exercises like the bench press, where similar biomechanical cues signal the onset of failure.

The rest of the report is structured as follows. Section 2 provides a high-level overview of the task, formulating the failure prediction problem and discussing parallels in existing literature. Section 3 details the novel dataset collected for this study. Section 4 describes the preprocessing pipeline, covering pose estimation, repetition segmentation, feature engineering, and data augmentation. Section 5 presents the model architectures, including our proposed Hierarchical LSTM and a linear regression baseline. Section 6 outlines the experimental setup, training procedures, and evaluation metrics. Section 7 presents the results comparing the models. Section 8 discusses and concludes the findings and their implications. Section 9 suggests directions for future work.

2. High-level task overview

2.1. Formulation

We formulate the prediction of muscular failure as a sequential learning task, focused on detecting the point at which a user is about to fail a repetition. Consider a single exercise set performed to failure, comprising a sequence of repetitions (reps), where each rep represents a discrete timestep. Let the final eight reps preceding failure be denoted as $[x_7, \dots, x_0]$, where x_0 is the last successfully completed rep before failure occurs, and x_7 is the eight-to-last rep in the sequence. Each x_t corresponds to pose-derived features (e.g., joint angles, velocities) for the t -th rep relative to failure. Conceptually, this is best understood as a boundary detection problem: the objective is to accurately identify the final safe rep x_0 , without issuing premature warnings.

Our aim is twofold:

1. Maximise the likelihood that the model identifies x_0 as the final safe rep (i.e., $P(x_0 = t_0)$).
2. Minimise the likelihood of premature alerts by suppress-

ing predictions for all preceding reps $[x_1, \dots, x_7]$. Formally, we aim to minimise $\sum_{t=1}^7 P(x_t = t_0)$, where $x_t = t_0$ implies that rep x_t is incorrectly classified as last rep before failure.

We restrict our attention to the final eight reps, rather than the entire set (which may contain many more), for both practical and modeling reasons. From a practical perspective, early reps are not relevant—trainees are well aware when they are far from failure and do not require predictive support at that stage. From a modeling standpoint, our dataset is relatively small, and including early, easily distinguishable reps could dilute the model’s capacity to learn the subtle distinctions between near-failure and actual failure, which is our central research question.

2.2. Parallels in literature

In Section 2.1, we formulated the task as a boundary detection problem, where the goal is to identify the final safe repetition before failure. While this framing is intuitive and aligns with the discrete nature of repetitions, the task can also be approached from a second, complementary perspective: remaining useful life (RUL) estimation. Together, these two views offer a more complete understanding of the problem.

In the boundary detection view, the objective is to classify which timestep corresponds to the final safe rep. This framing resembles a Hidden Markov Model (HMM), where hidden states (failure/non-failure) generate observable outputs (joint kinematics). However, standard HMMs are not suitable in this setting. Muscular failure is an absorbing state that occurs exactly once and cannot be reversed. These constraints break the Markovian assumptions of recurrence and state transition flexibility, and HMMs cannot enforce a single terminal event within a sequence.

Deep learning domains such as video action recognition or financial event forecasting also involve sequential prediction, but differ from our task in two key respects. First, they often lack discrete decision points—predictions may be made continuously or over sliding windows. In contrast, each rep in our setting defines a clear boundary at which a prediction must be made. Second, the events of interest in those domains typically occur multiple times or follow recurring patterns, whereas muscular failure is a one-time, irreversible event that terminates the sequence.

To address these structural constraints, we adopt an alternative framing based on Remaining Useful Life (RUL) estimation—a task in prognostics where the goal is to estimate how much time remains before a system fails. In typical RUL settings, models are trained on sequential sensor data collected during normal operation (e.g., from engines or turbines), and aim to forecast the number of cycles remaining before a critical failure. As illustrated in Figure 1, this perspective maps cleanly onto our setting: the start of the 8-rep sequence corresponds to the potential failure point (the onset of degradation), the final successful rep is the functional failure point, and the failed rep itself marks the complete failure point. Our objective becomes to estimate

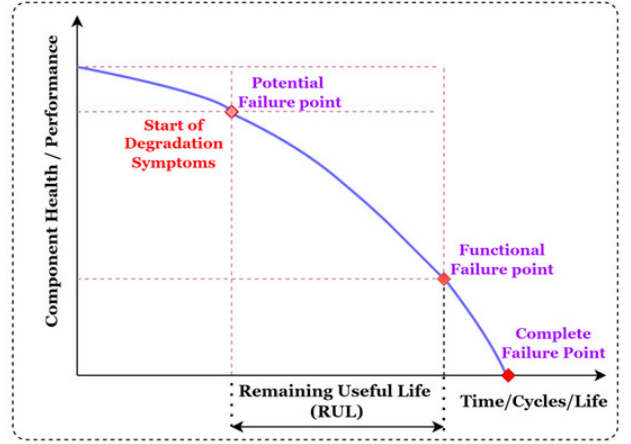


Figure 1. Concept of remaining useful life (RUL) from (Sayyad et al., 2023)

the number of reps remaining between potential and functional failure—that is, the RUL. We train the model using mean squared error (MSE), which encourages it to capture fine-grained fatigue dynamics across reps. Final predictions can then be thresholded to determine when to trigger an alert, e.g., when predicted RUL is 1.

This approach has two key advantages. First, it supports learning smooth temporal trends in biomechanical signals that might not be easily separable in a classification setting. Second, it aligns with the asymmetric risks of the task: failing to predict true failure has safety implications, while false positives reduce user trust. Compared to traditional RUL settings, however, which may model gradual degradation over thousands of cycles, our task involves much shorter sequences and with more abrupt degradation, so the model must be sensitive to rapid transitions while retaining context from earlier reps.

3. Dataset

Several fitness-related datasets exist, such as Fitness-AQA (Parmar et al., 2022), Countix (Dwibedi et al., 2020), and the Waseda Squat dataset (Matsui et al., 2019), addressing tasks like action quality assessment and repetition counting. However, none specifically capture exercise sequences performed until *muscular failure*. Existing datasets typically omit final repetitions where participants experience muscular failure, essential for our failure-prediction task.

To address this gap, we collected our own dataset. Participants performed bicep curls until muscular failure, recorded vertically from a strict profile angle to clearly capture shoulder, elbow, and wrist joints, ensuring minimal perspective distortion.

Data collection proceeded in two phases:

Pilot Collection: The researchers initially recorded approximately 20 sets to failure, experimenting with different clothing types, camera distances, and lighting conditions. It was found that dark clothing did not significantly impact

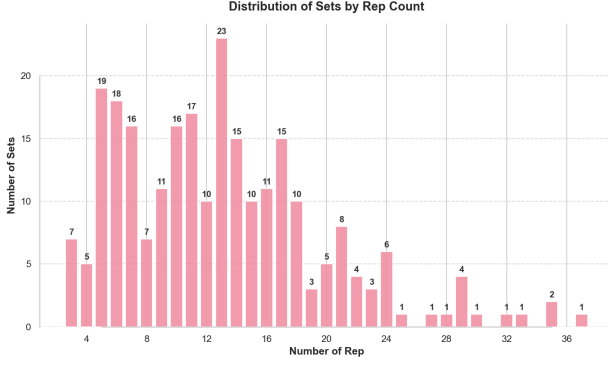


Figure 2. Distribution of repetition counts across the dataset. The dataset comprises 254 sets totalling 3272 repetitions (mean = 12.98, SD = 6.72).

pose estimation, provided shoulder contours were visible. However, baggy clothing noticeably reduced pose accuracy, leading us to instruct future participants to avoid loose clothing.

Participant Collection: Following ethics approval (application number 161250, approved on 2025-03-04), we conducted public data collection sessions at Appleton Tower Café. Volunteers performed bicep curl sets until muscular failure under consistent recording conditions.

All video sets were trimmed to include only complete repetitions, excluding idle movements before the first rep and any incomplete final reps. Sets showing severely compromised form, interruptions, or incomplete repetitions were removed. After trimming the videos to include only full, properly executed repetitions and removing sets with severe form issues, our final dataset comprised 254 unique sets from 66 participants, totalling 3272 reps. The mean number of repetitions per set was 12.98 with a standard deviation of 6.72 (see Figure 2 for the distribution). Participant information sheets and consent forms are provided in Appendices B and C, respectively.

4. Preprocessing pipeline

The following section covers the preprocessing pipeline for data. Figure 4 covers these steps.

4.1. Pose Estimation

We employed Mediapipe Pose to extract 2D joint coordinates (x , y) for the shoulder, elbow, and wrist joints nearest to the camera. To standardise across participants, the shoulder landmark position in the first frame was set as the origin (0, 0), and all other landmarks were translated accordingly. Pose estimation results were cached as .npz files containing normalised joint positions and timestamps.

4.2. Rep Segmentation

Videos were segmented into individual reps using a rule-based method based on elbow angles. The elbow angle

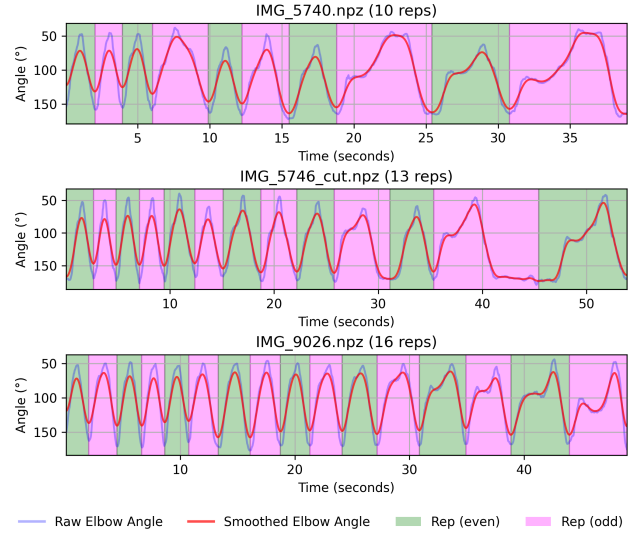


Figure 3. Three sets with smoothed elbow angle over time with automatic repetition segmentation. Alternating colours (green/pink) indicate detected reps. The top set displays irregular tempo and range of motion, the middle set includes a long rest period, and the bottom set shows consistently paced reps with increasing duration and mid-rep form dips in elbow angle.

was computed per frame using vectors connecting shoulder-elbow and elbow-wrist joints, then smoothed with a Gaussian filter to reduce noise. Repetitions were identified as segments between peaks in elbow angle, where peaks represent full arm extension.

Rest periods with minimal elbow angle change were not separately labelled and were included within segmented rep. Figure 3 illustrates examples of segmented elbow angle curves.

Segmentation parameters (smoothing factor σ , peak prominence, and minimum peak distance) were iteratively tuned through manual inspection to ensure accurate and consistent repetition identification. Only complete repetitions were included, excluding idle or incomplete movements.

4.3. Feature Engineering

Coming from our rep segmentation, each bicep curl rep is represented as a sequence of frames, where each frame contains 2D coordinates for the shoulder, elbow, and wrist joints. To transform these raw pose sequences into a compact, fixed-size input suitable for sequential modelling, we engineered a set of features that capture both spatial and temporal movement characteristics. We extracted seven features per frame, grouped into three categories:

Joint Position (4D): The elbow and wrist positions are expressed relative to the shoulder origin in the first frame, yielding four values per frame: elbow_x, elbow_y, wrist_x, wrist_y. This normalisation accounts for minor changes in camera framing and participant position, ensuring spatial consistency across samples.

Velocity (2D): For both the elbow and wrist, we measure

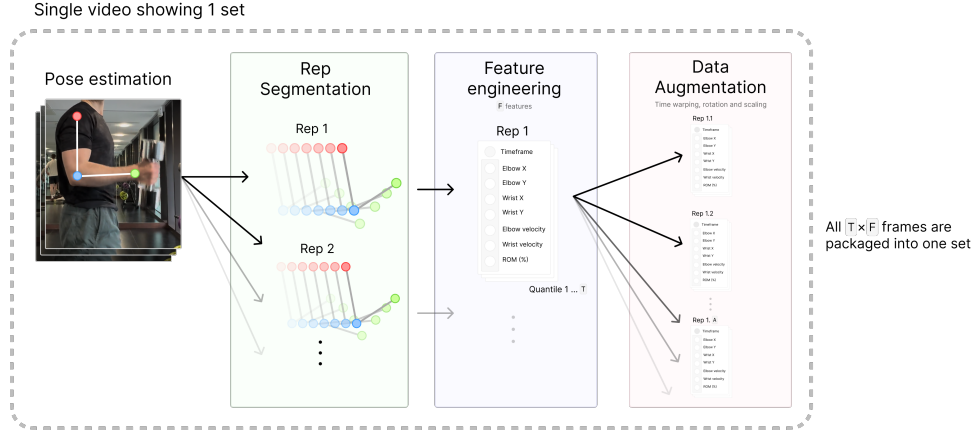


Figure 4. Full preprocessing pipeline. After raw video frames are pose-estimated, they are segmented into reps. Feature engineering extracts temporal and spatial characteristics (elbow/wrist positions, velocities and ROM). These features are then augmented through time warping, scaling, and rotation.

how quickly each joint moves between frames. These features capture movement dynamics such as slowing or hesitation, which are often observed in fatigued reps.

Range of Motion (ROM, 1D): We calculate the elbow angle in each frame by measuring the angle between the shoulder-elbow and elbow-wrist vectors. This gives a dynamic ROM signal that is then normalised per rep. A decreasing ROM across reps may indicate diminishing force output and loss of control, both precursors to failure.

To convert each variable-length rep into a fixed-size input, we sample values from the seven features at 20 evenly spaced quantiles. This avoids the need for zero padding, which can distort learning in sequence models. It also preserves temporal structure while producing a consistent shape for the LSTM. Each rep is represented as a 20×7 feature matrix, capturing key movement snapshots in a compact and consistent format.

4.4. Data Augmentation

To address the limited size of our dataset and improve model generalisation, we implemented a data augmentation strategy in line with established techniques in skeleton-based human action recognition (Xin et al., 2024). The goal was to introduce plausible variations in pose and motion without compromising biomechanical validity. We applied four main types of augmentation, each targeting different forms of variability:

Random Rotation: We applied random in-plane rotations of up to $\pm 5^\circ$ to all joint coordinates around the shoulder-centred origin. This simulates slight variations in camera angle or participant orientation. The rotation preserves limb proportions and relative distances, ensuring the integrity of motion patterns while increasing viewpoint diversity.

Scale Perturbation: Joint coordinates were randomly scaled by a factor between 0.95 and 1.05, uniformly in both axes. This accounts for natural differences in participant body size and positioning relative to the camera.

Such small scale variations have been shown to improve robustness without distorting kinematic features (Xin et al., 2024).

Temporal Warping: To reflect realistic variations in rep tempo, we implemented a time-warping procedure. Each rep’s joint sequence was resampled using a random factor between 0.85 and 1.15 and interpolated back to the original number of frames. This allows the model to learn temporal patterns independent of execution speed.

Additive Gaussian Noise: We added Gaussian noise with a standard deviation of $\sigma = 0.02$ to all joint coordinates, mimicking the natural jitter present in pose estimation systems.

All augmentations were applied consistently across full sets rather than individual reps. For each original set, we generated 2 augmented variants, effectively increasing the training set size fourfold. Importantly, augmented data was used exclusively during training. Validation and test evaluations were performed on unaltered, original sequences to maintain experimental integrity.

5. Model Architecture

5.1. Hierarchical LSTM

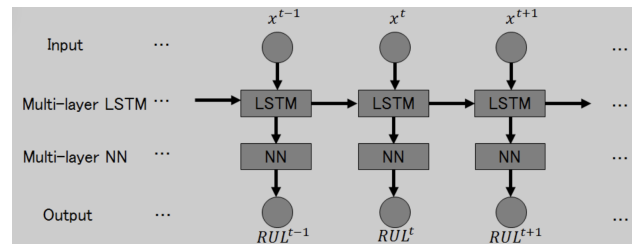


Figure 5. Unfolded LSTM structure for RUL estimation proposed in (Wu et al., 2018)

Modern RUL solutions favour recurrent neural networks, particularly LSTMs, over classical CNNs due to their abil-

ity to model variable-length sequences and long-term trends (Wu et al., 2024). Our architecture builds on a two-layer hierarchical LSTM framework similar to (Wu et al., 2018), but with key adaptations tailored to muscular failure prediction. In their setup (illustrated in Figure 5), sensor data from individual operational cycles is processed by a first LSTM layer, where each cycle’s hidden state propagates to the next cycle, and a neural network estimates RUL at each timestep.

We adopt a comparable hierarchical structure but reengineer it to reflect the unique temporal hierarchy of our setting. Here, our primary “timesteps” are reps, with each rep further subdivided into 20 timesteps of 7 ‘sensor’ features, resulting in two nested temporal scales.

This structure requires two modelling stages, shown in Figure 6. First, the timesteps within each rep are processed by an initial LSTM to generate a condensed representation of the rep. This output then feeds into a second LSTM that models dependencies across reps, analogous to how (Wu et al., 2018) model dependencies across operational cycles. The final RUL prediction synthesises outputs from both layers, allowing the model to disentangle intra-rep dynamics (e.g., muscle activation patterns within a rep) from inter-rep degradation trends (e.g., fatigue accumulation across reps).

To address the challenge of detecting failure within the short critical sequence (the last 8 reps before failure), we process sliding windows of N consecutive reps. The window size N was set to 3, based on experimental tuning. We found that a shorter window prioritizes abrupt biomechanical shifts—such as reduced elbow flexion velocity or range of motion—that often manifest suddenly in the final reps preceding failure. In contrast, larger windows risk diluting these signals with noise from earlier reps, which may lack predictive relevance. Sliding windows also address data limitations: with only 250 training sets, overlapping windows artificially expand the dataset (by generating sequences of reps 1–3, 2–4, etc.), mitigating overfitting and enabling robust learning of failure precursors.

5.2. Linear Regression Baseline

To establish a performance benchmark and validate whether our problem requires the temporal complexity of an LSTM architecture, we implemented a linear regression model.

The baseline model processes the same 3-rep sliding windows as our LSTM architecture, but instead of learning temporal representations directly, it relies on engineered features. These engineered temporal features are needed because linear regression lacks any built-in mechanism for capturing temporal relationships, unlike our LSTM architecture. For each window we compute summary statistics of joint coordinates (mean, standard deviation, minimum and maximum), kinematic variables from the most recent rep, differences between mean values of the latest rep and the preceding one to capture short-term fatigue trends, and cross-rep statistics that summarize patterns across the entire window.

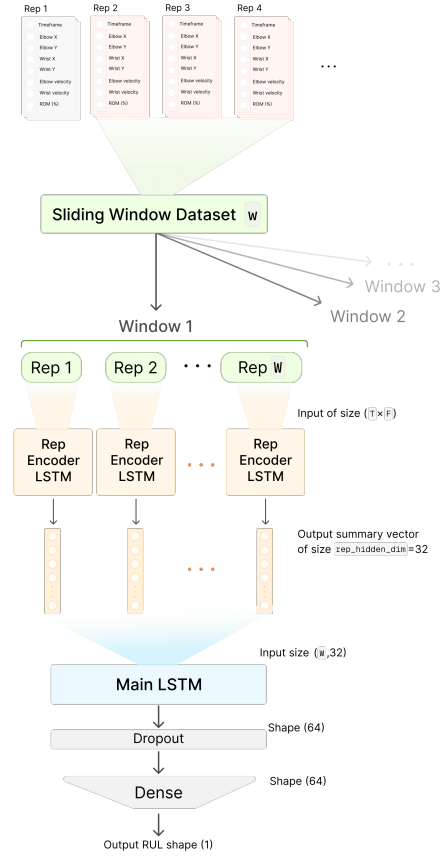


Figure 6. Our model architecture. Frame-level features (e.g. joint positions, velocities, range of motion) are encoded using a rep-level LSTM. The sequence of encoded reps is fed into a main LSTM, followed by dropout and a dense layer.

6. Experiments

6.1. Hierarchical LSTM

Our model employs two stacked LSTM layers within both the rep encoder and sequence modeling modules. Each LSTM layer contains 64 hidden units for rep-level feature encoding and 128 hidden units for inter-rep sequence modelling, with dropout ($p=0.5$) applied between layers to mitigate overfitting. Training runs for 75 epochs using the Adam optimizer (learning rate=0.005, weight decay= $1e-5$) with batch size 64. Learning rates are halved after 15 epochs of validation loss plateau via ReduceLROnPlateau scheduling.

Custom Loss function The loss function combines regres-

sion accuracy with asymmetric penalties:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \underbrace{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}_{\text{RMSE}} \\ & + \alpha \underbrace{\frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = 0 \wedge y_i > 0) |y_i|}_{\text{Premature alert penalty}} \\ & + \beta \underbrace{\frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i \neq 0 \wedge y_i = 0) |\hat{y}_i|}_{\text{Missed failure penalty}} \end{aligned}$$

The first term minimizes prediction error, the second penalizes premature failure alerts, and the third heavily penalizes missed failures. To determine appropriate weights for the asymmetric penalties, we fine-tuned the hyperparameters α and β through a grid search. Figure 7 visualizes the resulting R^2 surface across different combinations of α and β . We found that $\alpha = 1.0$ and $\beta = 0.8$ yielded the best performance, striking an effective balance between early warnings and missed detections.

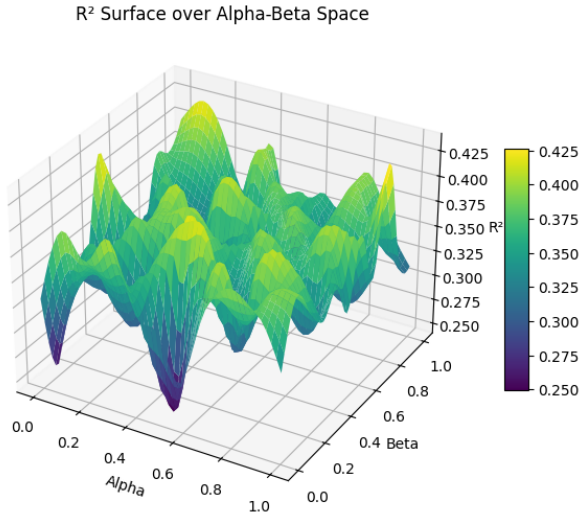


Figure 7. Experiments on alpha-beta

6.2. Linear Regression Baseline

We implemented a `scikit-learn`-based linear regression model to establish a performance benchmark against our LSTM architecture. The baseline uses the same feature engineering framework described for the hierarchical LSTM, but applies different training and prediction strategies.

The model operates on feature vectors of 224 dimensions created by applying temporal feature extraction to 3-rep sliding windows as described in Section 5.2. Each window produces statistics across multiple quantiles (20 per rep), capturing both intra-rep patterns and inter-rep changes. Unlike recurrent architectures, this approach explicitly encodes temporal relationships through engineered features.

For training, we used standard ordinary least squares regression without regularisation. We experimented with L1 and L2 regularisation variants but found they offered no significant performance improvement on our validation set. The model was trained on the same data split as our LSTM model, with identical augmentation procedures (2 augmented variants per original training set).

To ensure consistent evaluation across models, the continuous outputs from the linear regression were processed according to the methodology detailed in Section 6.3, including clipping to non-negative values and flooring to the nearest integer.

6.3. Evaluation Metrics

In addition to reporting continuous regression metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R^2 (aligned with our MSE training objective), we also convert the model’s outputs into a format suitable for binary classification, as motivated in the problem formulation section. Specifically, we take each predicted RUL value p and apply a flooring operation, mapping it to $\lfloor p \rfloor$. This transforms any prediction in the interval $[0, 1)$ into 0, which we interpret as imminent failure, meaning the current rep is likely the final safe one. This binning allows us to draw a clear decision boundary: a prediction of 0 means the model believes failure is imminent and an alert should be triggered, while higher values indicate the user is still safely within their rep range.

7. Results

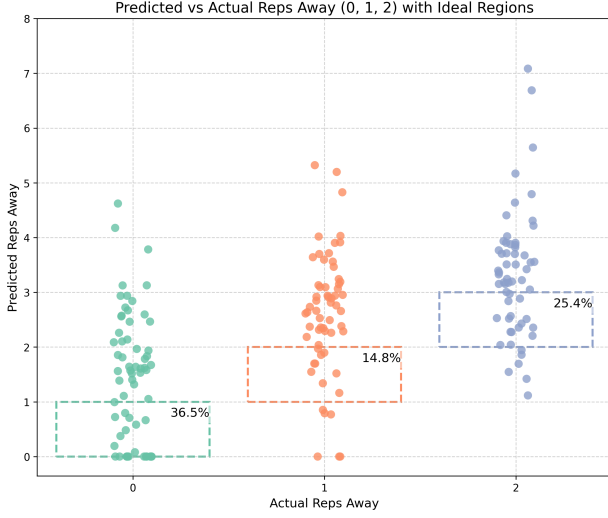


Figure 8. Linear regression results showing predicted vs. actual number of reps away from failure. Each dot represents a sample, with ideal prediction regions shown as dashed boxes. The percentage of predictions falling into each ideal region is also shown.

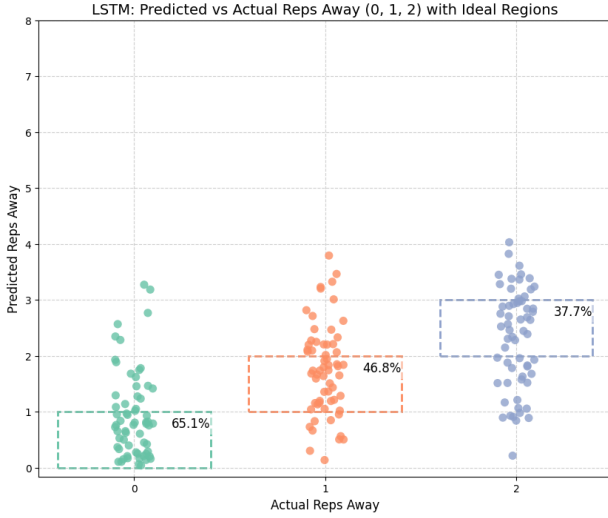


Figure 9. LSTM results showing predicted vs. actual number of reps away from failure. Each dot represents a sample, with ideal prediction regions shown as dashed boxes. The percentage of predictions falling into each ideal region is also shown.

The experimental results demonstrate that our Hierarchical LSTM significantly outperforms the linear regression baseline on the core task of final rep prediction, while maintaining competitive performance on overall sequence modelling. As shown in Table 1, the LSTM achieves 65.08% accuracy in identifying the final safe repetition (x_0), more than doubling the linear regression’s 36.52% performance. This substantial improvement validates our hypothesis that temporal dependencies and fatigue patterns in the final reps require nonlinear sequence modelling capabilities beyond what simple regression can capture. In addition to higher classification accuracy, the LSTM consistently achieves lower Mean Absolute Error (MAE) and Root Mean Squared

Model	Failure Rep			Overall			
	%correct	RMSE	MAE	%correct	RMSE	MAE	R^2
Linear Regression	36.52	1.55	1.14	85.77	2.05	1.56	0.37
Hierarchical LSTM	65.08	1.16	0.91	91.96	1.29	1.035	0.41

Table 1. Performance comparison between the Linear Regression baseline and the Hierarchical LSTM for rep failure prediction, where best results are highlighted in **bold**. The columns under “Failure Rep” report metrics when predicting the final safe repetition (rep x_0). The %correct in this section measures how often the model correctly identifies that final rep. Under “Overall”, metrics are computed across all predictions. Here, the Overall %correct represents the percentage of predictions for reps with a true remaining count higher than 1 that correctly avoid being erroneously labelled as failure (i.e. they do not trigger a premature failure alert).

Error (RMSE) across both failure rep prediction and overall sequence performance, indicating better calibration and more precise numerical predictions.

While the LSTM yields only a modest improvement in overall R^2 (0.41 vs. 0.37), this reflects our explicit design choice to prioritize boundary detection over global regression accuracy. As outlined in Section 2, the model’s core objective is identifying the final safe rep (x_0) – not minimizing error uniformly across all reps. Our asymmetric loss function ($\alpha = 1.0, \beta = 0.8$) codifies this priority by penalizing missed failures 25% more harshly than premature alerts, forcing the LSTM to focus capacity on modeling the final reps (x_0, x_1) where biomechanical failure signatures emerge.

R^2 , being a global goodness-of-fit metric, is thus less sensitive to performance at this boundary. We intentionally treat RUL regression as a proxy for detecting the failure boundary, rather than conventional sequence forecasting – a distinction clarified in Section 2.2. From this perspective, the LSTM’s 29% absolute gain in final-rep accuracy confirms it captures the degradation patterns preceding failure, while its stable R^2 shows it avoids overfitting noise in early reps irrelevant to safety decisions.

The linear baseline’s high overall %correct (85.77%) and low failure rep %correct (36.52%) reflect its risk-averse prediction strategy—by rarely predicting imminent failure, it minimises premature alerts but fails catastrophically when failure actually occurs. In contrast, the LSTM’s balanced performance (65.08% failure detection with 92% overall error rate) demonstrates its ability to navigate the asymmetric risk profile described in Section 6, providing more reliable failure warnings.

8. Discussion

Our findings provide strong evidence that joint movement patterns can be used to predict the onset of failure in strength training exercises. Both the linear regression and Hierarchical LSTM models demonstrated the ability to learn meaningful predictive patterns, with the LSTM significantly outperforming the baseline on the task of identifying the final safe repetition before failure.

As anticipated, the Hierarchical LSTM captured temporal dependencies more effectively than the linear model, learning richer representations of fatigue progression and exercise dynamics. It achieved a 29-point improvement in failure rep detection accuracy (65.08% vs. 36.52%) and maintained comparable performance in overall sequence modelling. This suggests that even with limited data, deep learning models can extract nuanced signals associated with physical exertion and imminent failure.

However, the LSTM's performance is not without limitations. Despite its gains, as illustrated in Figure 9 it failed to detect the failure rep (x_0) in approximately 35% of cases, and often triggered failure predictions prematurely—especially at x_1 , the penultimate repetition. This tendency toward early warnings, while arguably safer in real-time applications, may reduce the practical value of the system by limiting training efficiency or prematurely halting sets.

We attribute these limitations primarily to the size and granularity of our dataset. Distinguishing between the final and penultimate reps is a subtle task, requiring the model to learn very fine-grained temporal features. While we took measures to ensure that all sets reached genuine failure—including discarding ambiguous cases—some noise is inevitable. For instance, if a participant unknowingly had one more rep in reserve, the model would receive conflicting supervision. Such label noise disproportionately affects x_0 and x_1 , which are already underrepresented in the data distribution.

With a larger and more diverse dataset, we believe the model would be better equipped to capture the subtle degradation signatures unique to each individual's fatigue pattern.

9. Future Work

Our findings indicate that joint-only pose sequences are sufficient to predict muscular failure in bicep curls. A logical next step is to test whether this approach applies to higher-risk exercises like the bench press. While the bench press involves similar joint movements (primarily elbow extension) it poses a greater injury risk if failure goes undetected. Since our method relies solely on pose estimation from video, it could be particularly valuable in unsupervised settings, such as home gyms or public facilities without spotters.

Another promising direction is to explore transformer-based architectures for failure prediction. Transformers excel at capturing long-range dependencies in sequential data and have outperformed RNNs and LSTMs in tasks like action recognition and motion forecasting (Vaswani et al., 2017; Plizzari et al., 2021). In this context, transformers could better model subtle fatigue cues across reps using self-attention that are not being captured by our LSTM, while their ability to handle variable-length sequences could eliminate the need for manual windowing.

Failure detection could also benefit from multi-modal data,

combining pose data with signals like video frame features (e.g., facial strain), audio cues (e.g., breathing sounds), or wearable sensor data. While our current focus is on minimal-equipment deployment, a multimodal approach could improve accuracy for users willing to trade off convenience.

References

- Dwibedi, D. et al. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Lombardi, V. P. and Troxel, R. K. U.S. Deaths & Injuries Associated with Weight Training. *Medicine & Science in Sports & Exercise*, 35(5):S203, May 2003.
- Matsui, Y. et al. Waseda squat: Multi-angle pose estimation dataset for squat movement evaluation. In *ACM Multimedia*, 2019.
- Parmar, Paritosh, Gharat, Amol, and Rhodin, Helge. Domain knowledge-informed self-supervised representations for workout form assessment. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 105–123. Springer, 2022.
- Plizzari, Carlo, Cannici, Marco, and Matteucci, Matteo. Skeleton-based action recognition via spatial reasoning and temporal stack learning. *arXiv preprint arXiv:2104.03493*, 2021.
- Sayyad, Sameer, Kumar, Satish, Bongale, Arunkumar, Kotecha, Ketan, and Abraham, Ajith. Remaining useful life prediction of the milling cutting tool using time-frequency-based features and deep learning models. *Sensors*, 23(12):5659, 2023. doi: 10.3390/s23125659. URL <https://doi.org/10.3390/s23125659>.
- Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Wu, Fuhui, Wu, Qingbo, Tan, Yusong, and Xu, Xinghua. Remaining useful life prediction based on deep learning: A survey. *Sensors*, 24(11):3454, 2024. doi: 10.3390/s24113454.
- Wu, Yuting, Yuan, Mei, Dong, Shaopeng, Lin, Li, and Liu, Yingqi. Remaining useful life estimation of engineered systems using vanilla lstm neural networks. *Neurocomputing*, 2018.
- Xin, Chu, Kim, Seokhwan, Cho, Yongjoo, and Park, Kyoung Shin. Enhancing human action recognition with 3d skeleton data: A comprehensive study of deep learning and data augmentation. *Electronics*, 13(4):747, 2024. doi: 10.3390/electronics13040747.

Participant Information Sheet

Project title:	Predicting Rep Failure Using Pose-Based Deep Learning
Principal investigator:	Michal Kobiela
Researcher collecting data:	Tomas Maillo, Caterina Mammola
Funder (if applicable):	N/A

This study was certified according to the Informatics Research Ethics Process, reference number 161250. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The research team comprises Tomas Maillo (s2238874) and Caterina Mammola (s2185650) Computer Science students at the School of Informatics taking the 2025 Machine Learning Practical course. They are supervised by Michal Kobiela (m.kobiela@sms.ed.ac.uk) part of the Staff of Machine Learning Practical.

What is the purpose of the study?

The purpose of the study is to build a dataset of bicep curl exercise sequences, which will be used to train a deep learning model to predict muscular failure. This model aims to enhance safety during strength training by alerting users prior to reaching a point of failure, therefore reducing the risk of injury.

Why have I been asked to take part?

You have been asked to take part because you are within the targeted demographic for this study. We are recruiting individuals who have the ability to perform a series of bicep curls safely. Your participation will help us create and refine a dataset that is representative of real-world exercise conditions.



Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, up until we record your bicep curls without giving a reason. After this point, it will no longer be possible to withdraw because we are not collecting any data that would allow us to identify you.

What will happen if I decide to take part?

If you decide to take part, you will be recorded while performing bicep curl exercises in a gym or designated exercise area. The session will involve:

- Recording video footage of your unilateral (one-arm) bicep curl exercise for series of repetitions. We will then ask you to perform the same exercise with your other arm.
- You will not be asked to lift any weight that you are not comfortable with lifting.
- Data including joint coordinates derived from pose estimation will be extracted.
- The recording will last approximately 5 minutes and may include multiple sets as per our exercise protocol.
- There is no physical risk or discomfort beyond your usual exercise routine.

Are there any risks associated with taking part?

There are no significant risks associated with participation. The exercise session will involve performing standard strength training movements with weights that are not challenging, which you would normally undertake as part of your usual workout routine. Nonetheless, please follow any exercise guidelines to avoid injury.

Are there any benefits associated with taking part?

While there are no direct financial or material benefits to you from participating in this study, your involvement could contribute to research that may improve safety measures in strength training.



A. What data are you collecting about me?

B. Participant Information Sheet The research is completely anonymous: We are not collecting any information that could, in our assessment, allow anyone to identify you. Your signed participant consent form will be kept separately from your responses and destroyed in 6 months from signature.

We are collecting video recordings of you performing bicep curls (your face will not be in the shot), from which joint coordinate data will be extracted using pose estimation techniques. The data collected will be fully anonymised, and no personal identifying information will be linked to your exercise data.

What will happen to the results of this study?

Results of this study could be published online but this will not include your anonymised data.

The results will also be used as a submission to Coursework for the School of Informatics' course Machine Learning Practical.

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Tomas Maillo (s2238874@ed.ac.uk).

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be made available on <http://web.inf.ed.ac.uk/infweb/research/study-updates>.

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Tomas Maillo (s2238874@ed.ac.uk).



Participant number: _____

C. Participant Consent Form **Participant Consent Form**

Project title:	Predicting Rep Failure Using Pose-Based Deep Learning
Principal investigator (PI):	Michal Kobiela
Researcher:	Tomas Maillo, Caterina Mammola
PI contact details:	m.kobiela@sms.ed.ac.uk

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

1. I agree to being video recorded.

<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------

Yes No

2. I agree to take part in this study.

<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------

Yes No

Name of person giving consent

Date
dd/mm/yy

Signature

Name of person taking consent

Date
dd/mm/yy

Signature



THE UNIVERSITY of EDINBURGH
informatics