Deep Learning for Squat Phase Classification in Powerlifting: LSTM Approach VS Traditional ML Approach

Fort Hunter, Blake Milstead, Doga Topcicek, Linda Ungerboeck, Ryan Franqui
COSC 424/525
The University of Tennessee
Emails: {lhunte21, bmilstea, dtopcice, lungerbo, rfranqui}@vols.utk.edu

***Abstract*—We present a computer vision system for classifying squat movement phases in powerlifting using pose estimation and machine learning. Leveraging Mediapipe's Pose Model, we extract 3D skeletal data (33 keypoints at 0.3s intervals) from professional lifting videos. Our approach compares four classifiers—Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Random Forest, and XGBoost—to distinguish concentric (upward) from eccentric (downward) movements. The system prioritizes recall to minimize false negatives, essential for injury prevention. We address dataset challenges such as spotter interference using YOLO-based video preprocessing. Future work will extend the binary framework to detect specific form errors including spinal misalignment, knee valgus, and foot positioning. This system offers a low-cost, accessible alternative to professional coaching for technique correction in powerlifting.**

I. Introduction

The use of computer vision to analyze and correct athletic movements has been a topic of interest for researchers and practitioners alike, with applications ranging from injury prevention to performance enhancement. The importance of proper technique in powerlifting exercises cannot be overstated, as improper form can lead to serious injuries and long-term damage. Despite the growing recognition of this issue, many athletes struggle to identify weak points in their technique due to a lack of immediate feedback or access to expert coaching.

Current solutions for addressing these issues often rely on personal training sessions with experienced coaches, which can be expensive and inaccessible to those who cannot afford them. Furthermore, even trained professionals may find it challenging to pinpoint subtle errors in an athlete's form without extensive practice and experience. The need for a cost-effective solution that provides real-time feedback on posture correction has become increasingly pressing.

In response to this challenge, our team has developed a novel approach using computer vision techniques to analyze videos of powerlifting exercises. By estimating joint data from video analysis, we can provide immediate feedback on areas that require improvement, enabling athletes to adjust their technique without relying on external guidance. This technology offers a game-changing solution for both beginners and professionals seeking to optimize their training while minimizing the risk of injury.

The goal of this work is to construct an accurate classifier capable of distinguishing between concentric and eccentric movements in squat exercises. By achieving this objective, we aim to provide athletes with a valuable tool that can be integrated into their existing training regimens, promoting safer and more effective practice without breaking the bank. The rest of this paper outlines our technical approach, baseline selection, implementation, and future extensions of our network.

II. Literature Review

Human pose estimation plays an important role in analyzing athletic movements, particularly in the context of powerlifting, where proper technique is crucial for both performance optimization and injury prevention. Recent advancements in computer vision have significantly improved the ability to analyze human posture and movement during exercises such as squats.

One method for pose estimation is Mediapipe, a framework that utilizes machine learning models to extract 33 key joint landmarks from the video data. This framework enables real-time 3D skeletal representation, which can be utilized to assess and correct posture during exercises [1]. Mediapipe's Pose Model offers high accuracy and efficiency, making it ideal for applications in sports performance analysis and injury prevention.

In terms of classifying movements, previous research has focused on using machine learning models to distinguish between different phases of an exercise. For example, the use of random forests and gradient boosting classifiers has proven successful in analyzing dynamic movements in powerlifting exercises. These classifiers are particularly useful due to their ability to handle complex, nonlinear relationships inherent in human posture data [2]. Gradient boosting methods, in particular, offer strong performance for detecting subtle variations in joint movement, which is essential for accurately identifying concentric and eccentric phases during squat exercises.

Deep learning models, especially recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, have also shown promise in recognizing time-dependent activities such as squats. These models excel in learning from sequential data, making them ideal for tasks that require tracking movement over time [3]. Incorporating LSTMs into powerlifting

pose analysis has the potential to improve the accuracy of movement phase classification by leveraging temporal dependencies between key body points during the squat.

Furthermore, recent studies have explored the integration of skeletal data and video analysis techniques for automatic correction of exercise technique. These approaches could provide athletes with real-time feedback on their performance, ultimately improving training outcomes and reducing the risk of injury.

In summary, the use of advanced pose estimation and machine learning techniques holds great promise for optimizing powerlifting techniques and providing athletes with valuable feedback for improving their performance.

III. Dataset

To train and evaluate our squat phase classification system, we constructed a dataset from approximately 100 annotated 3D pose videos derived from footage of professional athletes and certified personal trainers. Our primary sources included the Olympic Powerlifting Championships, YouTube tutorials, and training content from fitness professionals. These sources were deliberately chosen to capture correct squat mechanics across a diverse range of individuals in terms of height, weight, and gender, thereby improving the generalizability of our model. Each video captures a single, properly executed squat repetition and was processed using Mediapipe's Pose Model, which extracts 33 skeletal keypoints (x, y, z) at 0.3-second intervals.

Due to the limitations of the machine learning models requiring a fixed input size, we resized our training images to 800 by 600 pixels. This transformation is applied to our training and testing data along with a video the model predicts. While the image is somewhat distorted, MediaPipe still can predict the 3D pose data needed for the machine learning models.

From the raw 3D coordinates, we computed four biomechanical angles critical to squat assessment: left knee, right knee, left hip, and torso lean. These angles were calculated using vector-based methods and expressed in radians. To reduce temporal noise without distorting movement dynamics, we applied Gaussian smoothing with sigma values between 1.5 and 2.0. For binary frame-level labeling, we calculated the average of the left and right knee angles and used the 50th percentile as a threshold: frames with higher average knee angles were labeled as UP (concentric phase), and those below were labeled DOWN (eccentric phase). The labeled frames were then segmented into overlapping sequences of 30 frames (approximately 9 seconds) for input to sequence-based models like LSTMs. Each sequence was assigned a label via majority vote across its frames to capture temporal patterns in squat movement.

All features were standardized using z-score normalization, with means and standard deviations stored for future inference. To address the class imbalance between UP and DOWN phases, we utilized a weighted random sampler during training to ensure balanced batch composition. One challenge we encountered involved tracking accuracy due to the presence of spotters assisting lifters during competitions. Mediapipe occasionally misidentified the spotter as the primary subject, especially when the spotter stood behind the lifter. This was combated with the use of YOLOv3 to track the most centered person in the video. This technique worked well but did not entirely solve the problem since some of the videos that we collected, mostly the Olympic videos, had a spotter in too close proximity for either model to differentiate.

The data was split into a 80% training, 10% testing, and 10% validation split. The smaller size of our dataset made training a more complicated model like a LSTM difficult. As one can see in our loss curves, the model does not learn much after the first couple of epochs. This provides insight into why our less complex models performed better just based on the size of our dataset. While the size of this dataset worked for a simple binary classification problem, in order to make it work for a multi-classification problem, that would be needed to provide real feedback, would require a much larger dataset.

To validate the quality of our extracted biomechanical features, we performed Principal Component Analysis (PCA) on the four angle-based features (left/right knee, left hip, torso lean). The resulting 2D projection (Figure I) shows clear structure in the data, with a distinguishable separation between the UP and DOWN phases. This confirms that the selected features encode meaningful variation and are suitable for phase classification.

Additionally, the overall label distribution (Figure II) demonstrates a near-even split between UP and DOWN frames across the dataset. This balance supports fair model training and reduces bias in performance metrics. Together, these visualizations confirm that our preprocessing pipeline produces structured, informative data ideal for both sequence and non-sequence models.

IV. Baseline Solution

To establish a robust foundation for our squat phase classification system, we explored multiple machine learning approaches to evaluate their effectiveness in distinguishing between eccentric (downward) and concentric (upward) motion during squats.

The dataset consisted of annotated video recordings from professional athletes and trainers, with each frame labeled as either UP or DOWN. We extracted 33 skeletal keypoints using MediaPipe, from which we computed four biomechanical features: left knee, right knee, left hip, and torso lean angles.

A. Traditional Machine Learning Models

We implemented two classic machine learning algorithms to serve as performance baselines:

Random Forest (RF) is an ensemble learning technique known for its robustness to noise and interpretability. Given the non-linear relationships inherent in joint-based movements, RF was selected for its ability to handle complex decision boundaries. Our RF model used 50 estimators, a maximum tree depth of 10, and default values for minimum samples at split and leaf nodes.

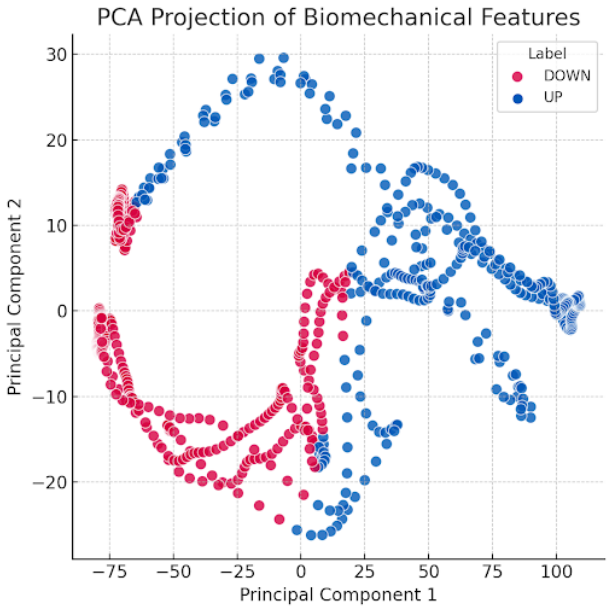

Fig. I: PCA projection of biomechanical features with UP/DOWN phase separation

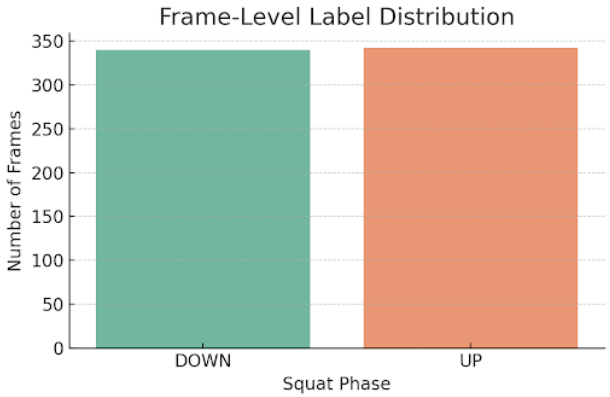

Fig. II: Class distribution of labeled squat frames (UP vs. DOWN)

Gradient Boosting (XGBoost) is a sequential boosting method that incrementally corrects its mistakes by combining weak learners. XGBoost was configured with 50 estimators, a learning rate of 0.01, a maximum depth of 3, and a gamma value of 0.1. These hyperparameters were chosen after grid search optimization.

Both models were trained using frame-level features and evaluated using accuracy, precision, recall, and F1-score, with a particular emphasis on recall to minimize false negatives—crucial in feedback systems aimed at injury prevention.

B. Deep Learning Models

To capture temporal dependencies and higher-order patterns, we developed two deep learning models:

Multi-Layer Perceptron (MLP) was implemented as a feedforward neural network. It served as a midpoint between traditional ML models and temporal models, capturing complex nonlinearities without modeling time.

- Input Features: 4 pose angles (left/right knee, hip, torso lean)
- Hidden Layers: [32, 16]
- Dropout: 0.4
- Batch Normalization: After each hidden layer
- Weight Initialization: Xavier
- Optimizer: Adam (learning rate = 1e-4, weight decay = 1e-5)
- Loss Function: Cross Entropy

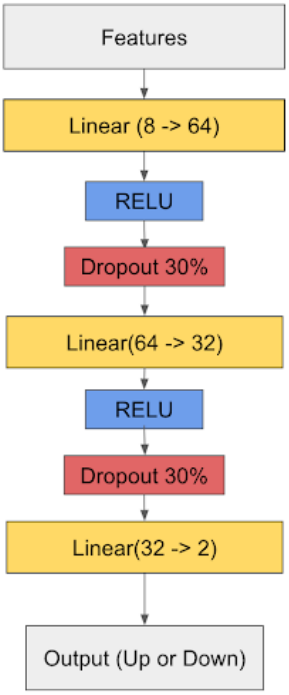

Fig. III: Architecture of the Multi-Layer Perceptron (MLP) model

Long Short-Term Memory (LSTM) was used to leverage the sequential nature of squat motion. It processes sequences of 30 frames, learning patterns across time rather than isolated instances.

- Input Features: 8 (4 pose angles + 4 motion deltas)
- Hidden Size: 256
- LSTM Layers: 2
- Sequence Length: 30
- Regularization: Dropout (implicit)
- Optimizer: Adam (learning rate = 1e-4)
- Loss Function: Weighted Focal Loss ($\gamma = 2$, $\alpha = 0.25$)

The LSTM model, while theoretically advantageous in modeling transitions, was more sensitive to noise in live pose estimation and had higher computational overhead compared to frame-based models.

C. Model Evaluation Focus

All models were assessed using accuracy, precision, recall, and F1-score. Given the goal of real-time athletic feedback, recall was prioritized to reduce the risk of missed detections in eccentric or concentric phases.

V. Results

We evaluated four models for binary squat phase classification using pose features extracted from video: a Multilayer Perceptron (MLP), Random Forest (RF), XGBoost (XGB), and a Long Short-Term Memory (LSTM) network. Each model was trained on knee, hip, and torso lean angles computed from skeletal landmarks using MediaPipe, with YOLOv3 employed for bounding box localization.

Among these models, the MLP achieved the highest test accuracy of 99.30% and an F1-score of 0.993. XGBoost and Random Forest followed closely, both achieving high precision and recall with minimal variance. While the LSTM model performed well on offline test data, its accuracy dropped significantly in real-world deployment due to sensitivity to pose estimation noise and smooth motion transitions.

TABLE I. Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
MLP	99.30%	0.993	0.993	0.993
XGBoost	98.9%	0.996	0.987	0.991
Random Forest	98.8%	0.996	0.982	0.988
LSTM	97.5%	0.975	0.976	0.975

A. MLP Loss Behavior

We trained two versions of the MLP model to evaluate the impact of regularization. The initial version showed classic signs of overfitting: as shown in Fig. IV, the training loss continually decreased while validation loss remained elevated and noisy, indicating poor generalization. This behavior suggested that the model was memorizing training data rather than learning robust patterns.

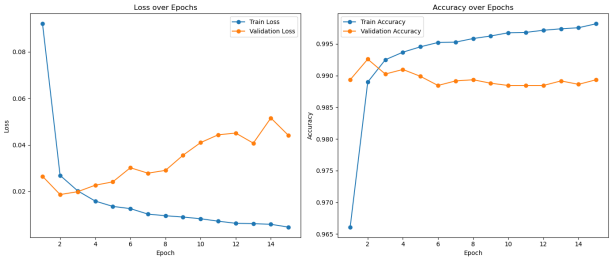

Fig. IV: Initial MLP training. Validation loss is consistently higher than training loss, indicating overfitting.

In the revised version, shown in Fig. V, we applied Xavier initialization, dropout layers, and additive Gaussian noise to the input. These regularization techniques increased the training difficulty, leading to slightly higher training loss, but significantly improved validation performance. Notably, the validation loss consistently remained lower than the training loss—an uncommon but interpretable result caused by the model's exposure to noise during training, making the validation set effectively easier.

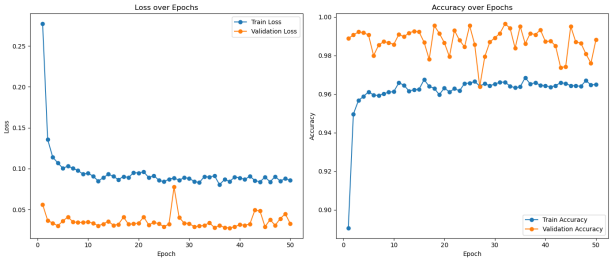

Fig. V: Improved MLP training. Regularization results in lower validation loss and better generalization.

B. Random Forest and XGBoost Tuning Trends

Unlike deep learning models, Random Forest and XGBoost do not optimize with continuous loss functions during training. Instead, their performance is tuned through hyperparameter searches. Figs. VI and VII show validation accuracy across parameter combinations. Both models demonstrated consistent performance and low variance, with accuracy stabilizing across various depth and estimator settings. These curves illustrate the models' robustness and relative insensitivity to moderate hyperparameter changes.

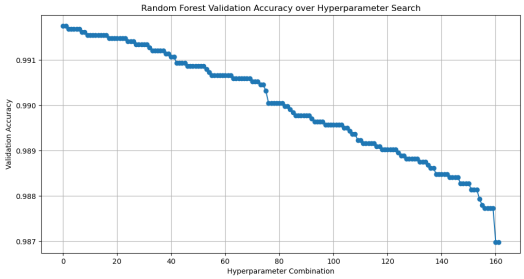

Fig. VI: Random Forest validation accuracy across hyperparameter combinations. The model maintains high accuracy with low variance.

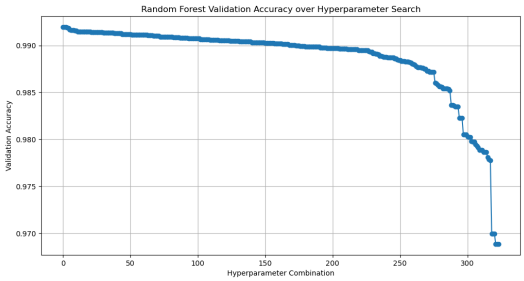

Fig. VII: XGBoost validation accuracy across hyperparameter combinations. Rapid convergence and consistency are observed.

C. LSTM Training Behavior

The LSTM model showed promising offline training behavior, as seen in Fig. VIII, where both training and validation loss decreased smoothly. However, this apparent convergence masked its poor generalization in real-time scenarios. The model struggled with sequential input degradation caused by pose jitter and smooth transitions, which made it less reliable when applied outside static datasets.

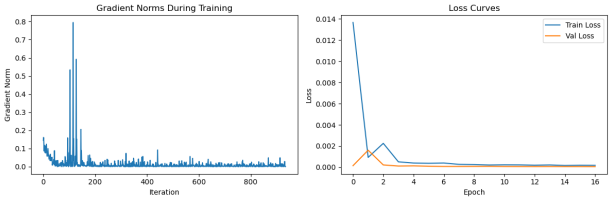

Fig. VIII: LSTM training and validation loss curves. Although performance is strong offline, generalization in real-time settings was limited.

D. Confusion Matrices and Model Comparison

Fig. IX presents the confusion matrices for the MLP, Random Forest, and XGBoost classifiers. All models demonstrated excellent class separation, with few misclassifications. The MLP had the highest true positive rate across both classes, while Random Forest and XGBoost showed strong agreement near phase transition boundaries.

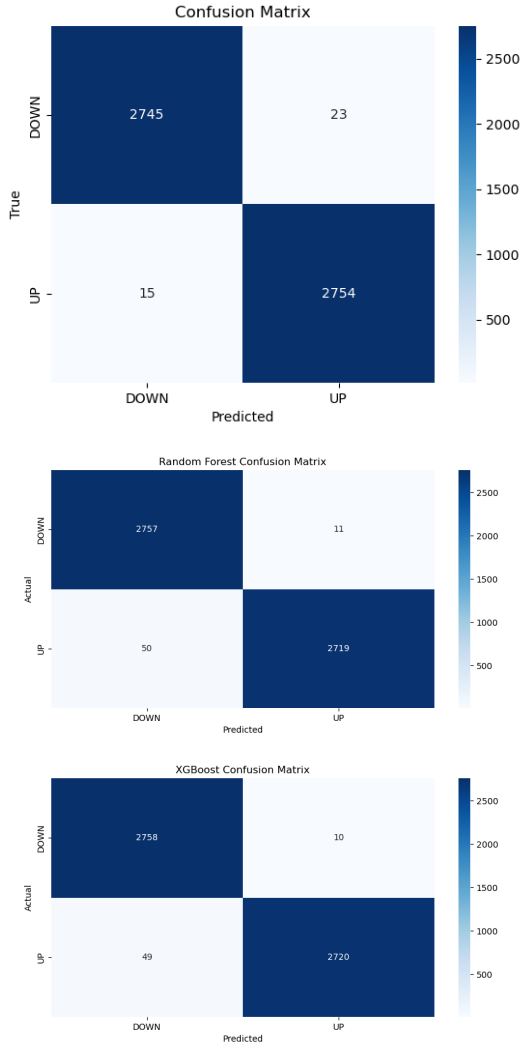

Fig. IX: Confusion matrices for MLP (top), Random Forest (middle), and XGBoost (bottom). All models show strong accuracy and minimal error.

VI. Conclusion

This work presents a real-time squat phase classification framework that combines YOLO-based person detection, MediaPipe pose estimation, and machine learning models trained on biomechanical features. Among the evaluated models, the regularized MLP achieved the highest test accuracy and F1-score, thanks to effective use of dropout, Xavier initialization, and input noise.

Random Forest and XGBoost also performed well, with hyperparameter tuning yielding stable high validation accuracy. Their resilience to overfitting and strong performance on frame-based features make them practical for real-world use. The LSTM, although effective in offline testing, struggled to generalize in real-time, highlighting the challenges of modeling temporal dependencies with noisy sequential data.

Visualizations of loss behavior and confusion matrices support these findings, illustrating that frame-based models, when

well-regularized and tuned, outperform sequence models under real-world constraints. Future work will focus on enhancing temporal model robustness through attention-based architectures and expanding the classification scheme to identify specific form deficiencies such as valgus collapse and improper spine alignment. Additional refinement of the YOLO cropping and pose estimation process may further improve consistency and reduce noise in model inputs.

VII. Proposed Extension

While the current model effectively distinguishes between concentric (upward) and eccentric (downward) squat phases, it provides only a binary assessment of correctness. This limits the system's ability to deliver actionable feedback to athletes. To address this, we propose expanding the classification framework to identify specific biomechanical deficiencies commonly observed in squat performance.

Future iterations of the model will incorporate multi-class labels that characterize particular faults during both upward and downward phases. In addition to the existing "correct" class, we will include three frequent error categories: (1) *non-neutral spine*, indicating loss of proper back alignment; (2) *valgus knee collapse*, referring to knees caving inward due to poor control or muscular imbalance; and (3) *excessive stance width*, which may compromise balance or depth. These labels will be phase-specific, resulting in four distinct classes per movement direction: *correct_up*, *non_neutral_spine*, *knees_caving*, and *feet_too_wide*.

Additionally, we aim to investigate and address the performance limitations observed in the LSTM model. While the LSTM performed well in controlled evaluations, it exhibited substantial degradation on real-world data, largely due to its sensitivity to noisy or inconsistent pose estimates. To improve its robustness, we will explore alternative sequence modeling strategies such as attention-based mechanisms and Transformer architectures, as well as enhanced preprocessing techniques to smooth temporal input sequences. Improving the temporal modeling capability of the system is critical for capturing subtle transition dynamics that frame-based models may miss.

Beyond model improvements, we will explore deployment pathways for real-time feedback using augmented reality overlays or integration with wearable sensors. These enhancements aim to create a comprehensive, interpretable, and accessible tool for squat performance assessment and athlete training.

VIII. Distribution of Work

While the project was a collaborative effort throughout the semester, each team member contributed distinct strengths during different phases of development. Ryan led the literature review and background research, facilitated communication with our mentor Dr. Santos, assisted with data collection, and helped coordinate the final presentation. Doga focused on implementing the baseline models and developing the LSTM architecture, in addition to contributing significantly to both the midterm and final reports. Linda supported baseline model development, participated in the data labeling process, and contributed to writing and editing both reports. Fort worked extensively on data collection and cleaning, collaborated on model evaluation and tuning, and contributed to both written reports and the presentation. Blake was responsible for data preprocessing and augmentation for the LSTM, developed the frame and sequence labeling strategy using joint angle calculations, and contributed heavily to both the technical implementation and the final report. All team members collaborated on preparing and delivering the final presentation and co-authored the final report to ensure technical clarity and cohesion.

AI Disclosure

The author(s) would like to acknowledge the use of **ChatGPT**, a language model developed by **OpenAI**, in the preparation of this assignment. ChatGPT was used for **brainstorming ideas, improving grammatical clarity, and refining technical phrasing in the Introduction and Literature Review sections**.

References

[1] J.-W. Kim, J.-Y. Choi, E.-J. Ha, and J.-H. Choi, "Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model," *Applied Sciences*, vol. 13, no. 4, p. 2700, Feb. 2023. DOI: 10.3390/app13042700.

[2] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and Random Forest Algorithms: An In-Depth Analysis," *Pakistan Journal of Scientific Research*, vol. 3, no. 1, pp. 26–31, 2023.

[3] Y. Yin, R. Yin, Y. Kim, and P. Panda, "Efficient Human Activity Recognition with Spatio-Temporal Spiking Neural Networks," *Frontiers in Neuroscience*, vol. 17, p. 1233037, 2023. DOI: 10.3389/fnins.2023.1233037.