


# CMPE 258 Final Project, Milestone 1

**Name of Leader:** Siddartha Kodaboina

**Other team members:** Preeti Prakash Aladakatti Rishabh Anand Kulkarni


## One sentence summary

Our Gen AI-based workout analyzer evaluates techniques in real-time, leveraging generative AI for personalized feedback, offering a more interactive approach than traditional solutions that focus solely on motion tracking or form correction using computer vision. 

## Proposed application and its novelty:

The proposed application, a Gen AI-powered Gym Workout Analyzer, is designed to provide personalized real-time feedback on workout form and technique, on exercises like push-ups, lunges, plank etc to enhance user safety and performance without the need for a personal trainer.

Prior related work:

In recent years, AI-driven fitness applications have advanced posture detection and real-time feedback; however, each method has its own limitations in practice. Ogundokun et al. [1] developed a posture detection method using transfer learning models (VGG16 and InceptionV3), enhanced by image augmentation and hyperparameter tuning  spite their effectiveness for detecting general objects, these pre-trained models have a difficult time detecting workout-specific postures, and their computational requirements make real-time implementation on mobile devices difficult. Duong-Trung et al. [2] proposed a real-time pose tracking system aimed at assisting users in self-guided fitness, employing pose estimation models to correct posture. However, environmental factors such as lighting reduce accuracy in detecting joint angles, especially for complex movements, while high processing requirements make mobile deployment challenging.

Erdaş and Güney [3] utilized wearable sensors with deep-learning to classify activities. While wearable data can capture general movements, it lacks the spatial and visual context essential for precise posture correction in real time. Additionally, noise from sensor data may reduce reliability during high-intensity exercises, limiting applicability for accurate form feedback. Lastly, Maji et al. [4] enhanced YOLO for multi-person pose estimation using object keypoint similarity loss, improving pose accuracy for multiple individuals. It, however is not practicable for single-user, real-time fitness applications on mobile platforms because of the increased computational complexity, especially in crowded places where overlapping poses can result in inaccuracies. They emphasize, the importance of lightweight, accurate models capable of

delivering personalized, real-time feedback without specialized hardware, while contributing to posture detection and activity recognition.

Our project approach, utilizing models like Modified ResNet50 and 3D CNN, offers key advantages by combining high accuracy with manageable computational requirements, allowing for real-time feedback on standard mobile devices. Contrasting from other general transfer learning or direct sensor data-based methods, our method accurately represents workout postures using frame-wise and temporal features, enhancing accuracy in form detection and correction. Apart from this, our system ensures personalized and actionable feedback by focusing on single-user scenarios without complex multi-person pose requirements, which makes our feedback responsive yet resource-efficient.

## Possible datasets:

Dataset/simulator name	Contents with examples	Pros	Cons
LSTM Exercise Classification: Push Up Videos	00 short (6 seconds) videos of push-ups, divided into 'Correct' and 'Incorrect' folders. Includes .npy files with body keypoints.	<ul style="list-style-type: none"><li>- Specifically designed for exercise classification</li><li>- Pre-processed with MediaPipe</li><li>- Consistent POV</li></ul>	<ul style="list-style-type: none"><li>- Limited to push-ups only</li><li>- Small dataset size</li><li>- Short video duration</li></ul>
Ha_500_v1_v1	Videos of various physical exercises, with about 20 videos per exercise	<ul style="list-style-type: none"><li>- Covers multiple exercise types</li><li>- Provides variety in movements</li></ul>	<ul style="list-style-type: none"><li>- Limited number of videos per exercise</li><li>- Potential inconsistency in video quality or format</li></ul>
Human3.6M	Large-scale dataset with 3.6 million 3D human poses and corresponding images	<ul style="list-style-type: none"><li>- Comprehensive 3D pose data</li><li>- Multiple actors and viewpoints</li><li>- Various everyday activities</li></ul>	<ul style="list-style-type: none"><li>- Not specific to workout exercises</li><li>- Large dataset size may require significant processing</li></ul>

# Possible models

Model name	Description with model diagrams/results	Classification acc. (top 1)	Pros	Cons	Notes
HOG + SVM	Utilizes Histogram of Oriented Gradients (HOG) features extracted from video frames, followed by a Support Vector Machine (SVM) classifier for posture classification. This method captures edge and gradient structures that are characteristic of human postures. <i>[Ref HOG]</i>	70%	Less computationally intensive; simpler implementation	Less effective in capturing complex spatial-temporal patterns	Traditional computer vision approach; serves as a baseline
3D Convolutional Neural Network (C3D)	Employs a 3D CNN that operates on sequences of video frames, learning spatial and temporal features simultaneously. The network consists of 3D convolutional layers followed by pooling and fully connected layers. <i>[Ref C3D]</i>	88%	Captures spatial-temporal features; good for video analysis	Requires substantial computational resources; longer training time	Effective for action recognition tasks in videos
Modified ResNet50	Fine-tuned ResNet50 model applied to individual frames extracted from videos. Additional layers and regularization techniques are added to enhance performance. The	94%	High accuracy; benefits from transfer learning	Processes frames individually, potentially missing temporal context	Implemented with added dropout and L2 regularization ; uses pre-trained weights

	model includes unfreezing the last 10 layers, adding dropout layers, and applying L2 regularization to prevent overfitting. [Ref ResNet]				
--	--	--	--	--	--

# Summary of Model Performance

- **Best Performing Model:** The **Modified ResNet50** achieves the highest classification accuracy at **94%**, indicating superior performance in identifying correct and incorrect workout postures.
- **Median Performance:** The **3D CNN (C3D)** model attains an accuracy of **88%**, effectively capturing both spatial and temporal features in videos.
- **Least Performing Model:** The **HOG + SVM** approach yields an accuracy of **70%**, reflecting limitations in handling the complexity of visual data and temporal dynamics in workout videos.

# Computational Considerations

- **Least Computationally Intensive:** The **HOG + SVM** model requires fewer computational resources and has a shorter training time, making it suitable for environments with limited hardware capabilities.
- **Most Computationally Intensive:** The **3D CNN (C3D)** demands significant computational power due to its 3D convolutional operations and processing of multiple frames simultaneously.
- **Modified ResNet50:** While computationally intensive due to its depth and fine-tuning process, it is less demanding than 3D CNNs in terms of processing temporal sequences.



# References

[1] R. O. Ogundokun, R. Maskeliūnas, and R. Damaševičius, "Human Posture Detection Using Image Augmentation and Hyperparameter-Optimized Transfer Learning Algorithms," *Applied Sciences*, vol. 12, no. 19, 2022, pp. 10156. doi:10.3390/app121910156.

[2] N. Duong-Trung, H. Kotte, and M. Kravčík, "Augmented intelligence in tutoring systems: A case study in real-time pose tracking to enhance the self-learning of fitness exercises," in *Proc. 18th European Conf. Technology Enhanced Learning (ECTEL)*, 2023.

[3] Ç. B. Erdaş and S. Güney, "Human activity recognition by using different deep learning approaches for wearable sensors," *Neural Process. Lett.*, vol. 53, 2021, pp. 1795–1809.

[4] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "Yolo-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2637–2646.

- **[HOG]**: Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1, 886–893 vol. 1. <https://doi.org/10.1109/CVPR.2005.177>.
- **[C3D]**: Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, 4489–4497. <https://doi.org/10.1109/ICCV.2015.510>.
- **[ResNet]**: Browne, R. F. (1997). 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings, June 21–23, 1994, Seattle, Washington. *Computers and Electronics in Agriculture*, 18(1), 67–69. [https://doi.org/10.1016/S0168-1699\(97\)01320-3](https://doi.org/10.1016/S0168-1699(97)01320-3).