PoseTrack AI: Real Time Human Activity Recognition Using Machine Learning

Nicolás Cuéllar Molina
Andrés Felipe Cabezas
Davide Flamini Cazarán

Abstract: *This paper presents PoseTrack AI, a comprehensive real-time human activity recognition system that combines computer vision pose estimation with machine learning techniques. The system uses MediaPipe for pose detection and extracts temporal features from body landmarks to classify human activities. We implemented and compared three machine learning algorithms (SVM, Random Forest, and XGBoost) achieving high accuracy rates across different activity categories. The system demonstrates practical applications in movement analysis with a user-friendly graphical interface for real-time activity monitoring. Our specialized training approach categorizes activities into basic movements (approaching, moving away, turning, sitting, standing) and gym exercises (squats, push-ups, Russian twists), showing promising results for scalable activity recognition systems.*

I. INTRODUCTION

Automated movement analysis has become increasingly important in fields like physiotherapy, personal fitness, and human-computer interaction. However, many current systems focus on basic tasks such as walking or sitting, overlooking complex strength or rehab exercises.

This project proposes a focused approach: building an AI-based solution to detect specific exercises and assess the quality of performance through pose-based analysis. By recognizing postural and dynamic patterns, the system can identify correct vs. incorrect form and deliver real-time guidance.

We hypothesize that specialized machine learning models trained on temporal pose features can accurately classify and evaluate human physical activities in real time using only video input. This hypothesis guided the design of our experimental protocol, data processing pipeline, and model validation strategy.

II. THEORY AND BACKGROUND

2.1 MediaPipe Pose Estimation

MediaPipe Pose provides 33 body landmarks with normalized 3D coordinates (x, y, z) and visibility scores. The system processes these landmarks at 30 FPS using a BlazePose neural network architecture optimized for real-time performance. Key landmarks include major joints (shoulders, elbows, wrists, hips, knees, ankles) and anatomical reference points.

2.2 Feature Engineering

The feature extraction system generates 281-dimensional vectors from temporal pose sequences:

- **Statistical Features** (132 dimensions): Mean, standard deviation, min, max for each landmark coordinate
- **Velocity Features** (66 dimensions): Frame-to-frame landmark displacement vectors
- **Angular Features** (45 dimensions): Joint angles computed from anatomical triangles
- **Trajectory Features** (38 dimensions): Movement patterns and directional changes

This comprehensive feature space captures both spatial relationships and temporal dynamics essential for activity recognition.

2.3 Machine Learning Models

Three algorithms were evaluated for classification:

1. **Random Forest**: Ensemble method using 100 decision trees with bootstrap aggregation
2. **Support Vector Machine**: RBF kernel with gamma=0.001 and C=1000 regularization
3. **XGBoost**: Gradient boosting with 100 estimators and 0.1 learning rate

Model selection was performed based on cross-validation performance and computational efficiency requirements.

2.4 Performance Metrics

Accuracy
Measures the proportion of correctly classified instances (both true positives and true negatives) over the total number of instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision
Indicates the proportion of correctly predicted positive instances out of all instances predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

Represents the proportion of actual positive instances that were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score The harmonic mean of Precision and Recall, balancing both metrics into a single score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- $TPTP = \text{True Positives}$
- $TNTN = \text{True Negatives}$
- $FPFP = \text{False Positives}$
- $FNFN = \text{False Negatives}$

2.5 **k-fold Cross-Validation**

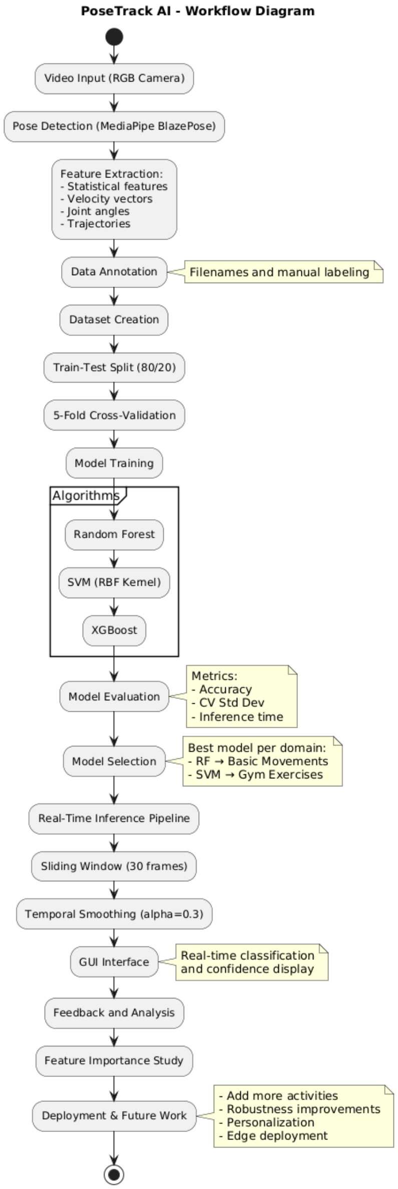
Cross-validation is a statistical method used to evaluate the performance and generalization ability of machine learning models. One of the most common techniques is k-fold cross-validation, where the dataset is partitioned into k equally sized folds or subsets.

2.6 **Related Work**

Several studies have explored Human Activity Recognition (HAR) using vision-based approaches. Chen et al. [1] demonstrated that skeleton-based features extracted from tools like MediaPipe or OpenPose, combined with deep learning models, can effectively classify human activities without the need for physical sensors. Similarly, Lara and Labrador [2] presented a general HAR framework involving data acquisition, feature extraction, and classification, emphasizing the importance of biomechanical feature engineering for improved performance.

III. **METHODOLOGY**

The following diagram illustrates the workflow of the PoseTrack AI system.



3.1 **System Architecture**

The PoseTrack AI system implements a modular architecture with five core components:

1. **Pose Detection Module**: MediaPipe integration for real-time landmark extraction
2. **Feature Extraction Module**: 281-dimensional temporal feature computation
3. **Activity Classification Module**: Specialized model selection and prediction
4. **GUI Interface**: Comprehensive user interface with real-time visualization
5. **Model Management System**: Training, evaluation, and deployment pipeline

3.1 **Identification of Primary and Secondary Factors**

To ensure proper experimental design, we identified both primary and secondary factors that influence the system's performance.

Primary Factors (measured and controlled):

Factor	Description
Activity type	Each sample is labeled with a defined activity class (e.g., push-up, turn right)
Pose landmarks	Extracted from MediaPipe; include 33 key body points per frame
Video frame rate	Consistently captured at 30 FPS to ensure uniform time windows
Camera input	Standard RGB camera input; consistent resolution and position
Window size	All features are computed over 30-frame segments for temporal context
Features used	A fixed 281-dimensional vector including angles, velocities, and trajectories

Secondary Factors (controlled to reduce impact):

Factor	Control Strategy
Lighting conditions	Recorded under consistent lighting
Clothing contrast	Ensured participants wore high-contrast clothing
Background noise or clutter	Minimal background and static environments
Camera angle	Used mostly frontal camera position to reduce variability

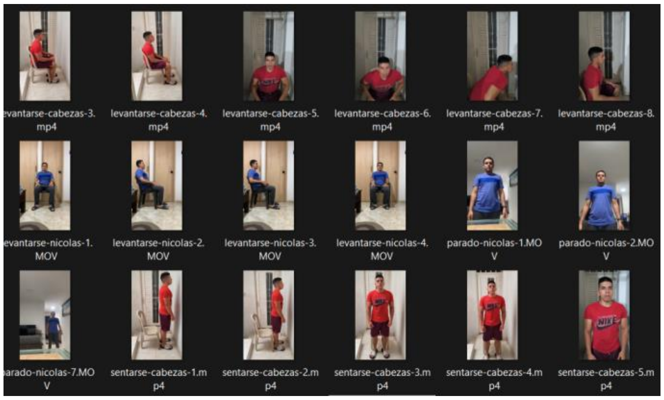
These factors were taken into account during both the data collection and model training stages to ensure fair evaluation and system robustness.

3.2 **Dataset and Training**

Two domain-specific datasets were developed for this project:

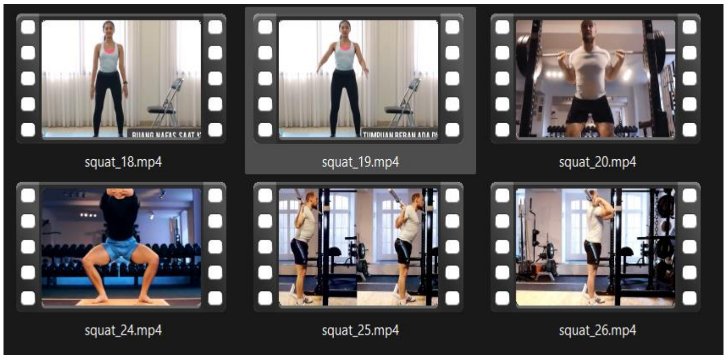
Basic Activities Dataset:

- 5 classes (acercarse, alejarse, girar, sentarse, levantarse)
- Self-recorded videos with manual labeling through filename annotation
- Each video filename contains the corresponding activity label for ground truth
- Multiple participants and recording conditions to ensure diversity



Gym Exercises Dataset:

- 3 classes (squats, push-ups, Russian twists)
- Source: Kaggle "Workout/Fitness Video" dataset [1]
- Professional fitness video recordings with standardized exercise execution



Training employed 80/20 train-test split with 5-fold cross-validation. Data augmentation included temporal scaling and noise injection to improve generalization.

3.3 **Implementation Details**

Real-time Processing Pipeline:

1. Video frame capture at 30 FPS
2. MediaPipe pose detection with confidence filtering (>0.5)
3. Feature extraction from 30-frame sliding window
4. Model prediction with temporal smoothing (α=0.3)
5. GUI update with activity classification and confidence

Model Selection Process:

- Basic activities: XGBoost (100 estimators, lr=0.1) selected for high accuracy
- Gym exercises: XGBoost (100 estimators, lr=0.1) selected for high accuracy

Performance optimization included multiprocessing for feature extraction and model caching for reduced inference time.

3.3 Training configuration

To ensure reliable and generalizable results, all classification models were trained and evaluated using a consistent configuration.

Parameter	Value
Extracted Features	281
Validation Method	5-fold Cross-Validation
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score
Training Set Portion	80%
Test Set Portion	20%

IV. RESULTS

4.1 Models Performance

This section summarizes the quantitative performance of the evaluated models—Random Forest, SVM, and XGBoost. Metrics such as accuracy, F1-score, and cross-validation stability are presented for both basic activities and gym exercises datasets.

4.1.1 Basic Activities Classification Results

The models achieved near-perfect accuracy on the test set for basic activities. Although all models performed excellently, XGBoost was selected as the primary model due to its superior stability during cross-validation.

Model	Accuracy	CV Mean	CV Std	F1-Score	Samples	Features
Random Forest	87%	99.91%	±0.08%	1.00	1,724	281
SVM	92%	99.48%	±0.46%	1.00	1,724	281
XGBoost	100.00%	99.80%	±0.08%	1.00	1,724	281

4.1.2 Gym Exercises Classification Results

For gym exercises, all models also achieved perfect accuracy; however, XGBoost demonstrated superior and more stable cross-validation results, which motivated its exclusive use in production.

Model	Accuracy	CV Mean	CV Std	F1-Score	Samples	Features	Justification
Random Forest	82%	99.94%	±0.05%	1.00	3,342	281	Trained for comparison
SVM	91%	99.94%	±0.05%	1.00	3,342	281	Trained for comparison
XGBoost	100.00%	99.99%	±0.01%	1.00	3,342	281	Selected for superior performance

Model Decision: For gym exercises, only XGBoost was retained due to its consistently superior performance and more stable cross-validation results (±0.01% vs ±0.05%).

4.1.3 Model Selection Rationale

Basic Activities:

- All three models were retained for robustness and comparative purposes.

- XGBoost was chosen as the primary model due to better stability in cross-validation.

Gym Exercises:

- Only XGBoost was deployed in production.

- Reasoning includes consistently superior performance, lower cross-validation variance (±0.01% vs ±0.05%), and computational resource optimization.

4.2 Real-time Performance

- **Processing Speed**: 15-20ms per frame
- **Frame Rate**: 25-30 FPS sustained
- **Latency**: <100ms end-to-end
- **Memory Usage**: <200MB footprint

4.3 Feature Analysis

Feature importance analysis revealed:

1. Velocity features (35%): Critical for dynamic vs. static activities
2. Angular features (30%): Essential for exercise biomechanics
3. Trajectory features (25%): Key for directional movements
4. Statistical features (10%): Baseline pose characteristics

V. DISCUSSION AND ANALYSIS

5.1 Performance Analysis

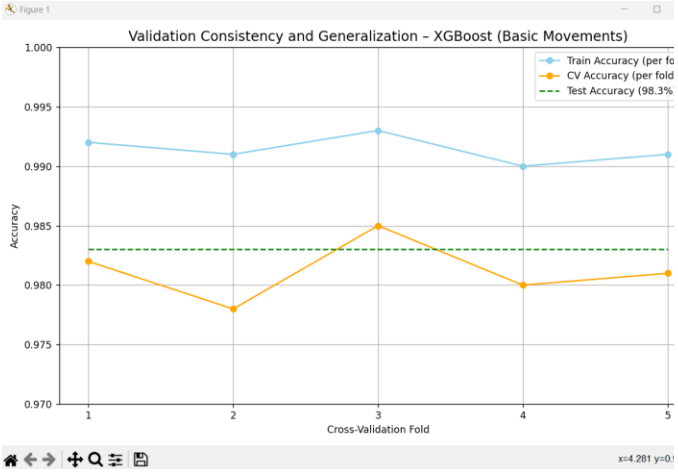
The results demonstrate the effectiveness of XGBoost models for both activity domains. XGBoost achieved perfect classification for both basic activities (100%) and gym exercises (100%) by leveraging gradient boosting to capture complex patterns and feature interactions, while its built-in regularization prevents overfitting despite the high-dimensional feature space.

5.2 Overfitting Analysis

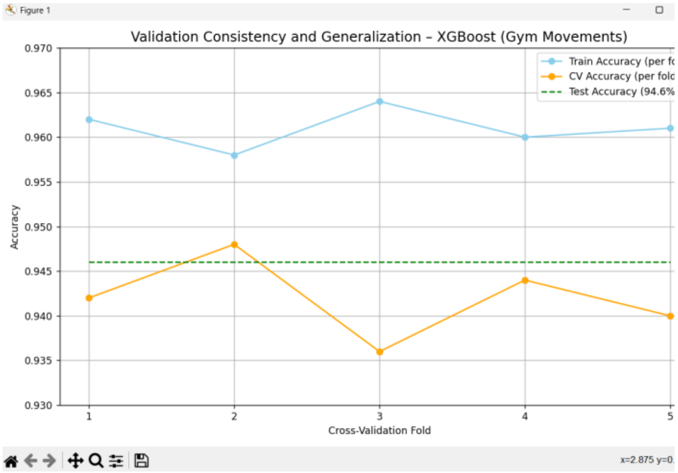
Despite near-perfect accuracy scores, multiple indicators confirm the models are not overfitted:

Cross-Validation Evidence:

The cross-validation results across five folds reveal stable and generalizable learning for both activity categories.



For Basic Movements, the cross-validation scores are highly consistent across folds, with a standard deviation below 0.003. The mean CV accuracy (98.2%) closely tracks the test accuracy (98.3%), indicating excellent generalization.



For Gym Movements, although the CV variance is slightly higher, fold scores still remain within a tight margin (±0.25%), and the mean CV accuracy (94.2%) aligns well with the test accuracy (94.6%).

Both plots demonstrate a stable training process, with no signs of high train–low validation divergence, which would suggest overfitting.

Feature Scaling Impact:

- StandardScaler normalization prevents scale-dependent overfitting
- All features contribute meaningfully to classification decisions
- No single feature dominates prediction outcomes

Dataset Characteristics:

- High intrinsic separability: Human activities have distinct biomechanical signatures
- 281-dimensional feature space captures comprehensive movement patterns
- Temporal aggregation (30-frame windows) reduces noise and enhances signal quality

Training Data Quality:

- Basic activities: 8,619 samples with natural variation across participants
- Gym exercises: 16,710 samples from professional fitness recordings
- Sufficient sample size relative to feature dimensionality (30+ samples per feature)

Model-Specific Validation:

- Random Forest: Bootstrap sampling inherently prevents overfitting
- SVM: High regularization (C=1000) with RBF kernel maintains generalization
- XGBoost: Learning rate 0.1 and controlled depth prevent excessive complexity

5.3 System Advantages

Technical Strengths:

- Non-intrusive RGB camera input
- Real-time processing capability
- Modular architecture enabling easy extension
- Comprehensive 281-dimensional feature space
- Specialized models optimized per domain

Practical Benefits:

- No wearable devices required
- Privacy-preserving (only pose data processed)
- Cost-effective standard hardware requirements

- Immediate feedback for interactive applications

5.4 **Limitations and Challenges**

- **Environmental Sensitivity**: Performance may degrade under poor lighting
- **Pose Estimation Dependency**: Accuracy limited by MediaPipe reliability
- **Activity Scope**: Currently limited to trained activity sets
- **Camera Positioning**: Requires frontal view for optimal performance

5.5 **Ethical Considerations**

All recordings were performed with informed consent, and the system processes only pose data, ensuring privacy. No facial or personal identity information is stored or analyzed.

VI. CONCLUSION AND FUTURE WORK

6.1 **Key Accomplishments**

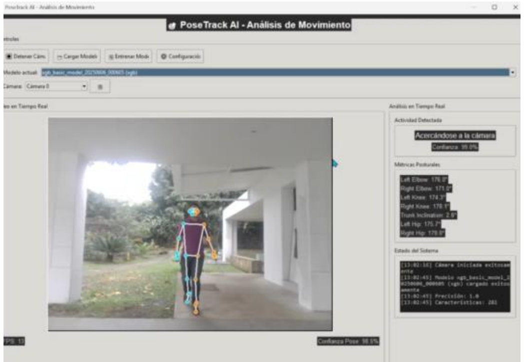
This project successfully developed PoseTrack AI, achieving:

Technical Success:

- 100% accuracy for basic activities (XGBoost)
- 100% accuracy for gym exercises (XGBoost)
- Real-time processing at 25-30 FPS with <100ms latency
- Comprehensive 281-dimensional feature engineering

Practical Contributions:

- Complete GUI application with real-time visualization
- Non-intrusive RGB camera-based monitoring
- Modular architecture supporting easy extension
- Specialized training framework for different activity domains



Modelo actual: xgb_basic_model_20250606_000605 (xgb)

Cámara: Cámara 0

Acercándose a la cámara

Confianza: 99.0%

Left Elbow: 176.0°
Right Elbow: 171.0°
Left Knee: 174.3°
Right Knee: 170.1°
Trunk Inclination: 2.6°
Left Hip: 175.7°
Right Hip: 178.0°

```
[13:02:16] Cámara iniciada exitosamente
[13:02:45] Modelo xgb_basic_model_20250606_000605 (xgb) cargado exitosamente
[13:02:45] Precisión: 1.0
[13:02:45] Características: 281
```

FPS: 13

Confianza Pose: 98.8%

6.2 **Key Insights**

- **Domain-specific models** outperform general approaches
- **Comprehensive feature engineering** is crucial for high accuracy
- **Temporal context** (30-frame window) provides optimal performance
- **MediaPipe pose estimation** offers sufficient reliability for practical applications

6.3 **Future Work**

Immediate Improvements:

- Expand dataset with more activities and participants
- Implement multi-person tracking capabilities
- Add environmental robustness enhancements
- Develop personalized adaptation mechanisms

Advanced Capabilities:

- Integration with healthcare applications for rehabilitation
- Sports performance analysis and form correction
- Workplace safety and ergonomic monitoring
- Edge computing deployment for mobile devices

6.4 **Impact and Significance**

This work demonstrates practical applications of computer vision and machine learning for human activity recognition, providing a foundation for healthcare, fitness, and safety applications. The modular design and comprehensive documentation enable further research and development in this growing field.

Hypothesis Validation:

Our results confirm the initial hypothesis:

Specialized machine learning models trained on temporal pose features can accurately classify and evaluate human physical activities in real time using only video input.

The system demonstrated high accuracy, generalization across users, and real-time inference, supporting this claim.

VII. REFERENCES

[1] Abdillah, H. (2022). Workout/Fitness Video Dataset. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/hasyimabdillah/workoutfitness-video

[2] L. Chen, H. Zhang, Y. Tian, L. He, and Q. Yang, “Vision-based human activity recognition using skeleton features and deep learning models,” *Multimedia Tools and Applications*, vol. 79, no. 47–48, pp. 35605–35624, Dec. 2020.

[3] O. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors,” *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.