EFFICIENT POSE-BASED HUMAN ACTIVITY RECOGNITION THROUGH SKELETAL FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION

Thesis Submitted
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
SOFTWARE ENGINEERING

by

Aman Yadav
(2K24/SWE/12)

Under the supervision of

Dr. Abhilasha Sharma

Associate Professor, Software Engineering

Delhi Technological University, Delhi, India

To the
Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

July 7, 2026

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

ACKNOWLEDGEMENTS

I express my deepest and humble thanks to my supervisor Dr. Abhilasha Sharma (Research supervisor) for providing with guidance, constant help and invaluable feedbacks through the whole period of the research. Without her advice, help, support, experience and patience, this project would not have achieved its present form.

I would like to convey my heartfelt thanks to Dr. Roshni Singh for all her encouragement, guidance, and valuable suggestions all throughout the research work. Her experience and encouragement supported me a lot in completion of the project.

I am very thankful to the Department of Software Engineering, Delhi Technological University(DTU) for the well-developed academic and necessary resources for undertaking this project and the faculty members for their guidance and imparting valuable knowledge to me.

I am grateful to my friends and peers for their endless encourage ments, interesting talks and help which make the work quite interesting and easier for me.

I am also thankful to my parents Mrs. Sanju Yadav and Mr. Birendra Kumar Yadav for the endless help, moral support and encouragement they provided to me all through the process of the completion of this research thesis. Their faith and love paved a way for the completion of my project.

I thank one and all, who has been a support in completing this thesis.

Aman Yadav

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Aman Yadav, hereby certify that the work which is being presented in the thesis entitled **"Efficient Pose-Based Human Activity Recognition Through Skeletal Feature Engineering and Dimensionality Reduction"** in partial fulfillment of the requirements for the award of the Degree of Master of Technology submitted in the Department of Software Engineering, Delhi Technological University, is an authentic record of my work carried out during the period from August 2024 to May 2026 under the supervision of Dr. Abhilasha Sharma.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Aman Yadav

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor | Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that Aman Yadav (24/SWE/12) has carried out his research work presented in this thesis entitled "Efficient Pose-Based Human Activity Recognition Through Skeletal Feature Engineering and Dimensionality Reduction" for the award of Master of Technology from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Abhilasha Sharma
Associate Professor
Software Engineering
Delhi Technological University
Delhi, India

Date: July 7, 2026
Place: Delhi

Abstract

Human Activity Recognition (HAR) plays a vital role in applications such as fitness monitoring, rehabilitation, surveillance, and intelligent human-computer interaction systems. While recent deep learning approaches based on RGB video data have achieved high accuracy, they are computationally expensive, require large-scale datasets, and are sensitive to variations in background and illumination. To address these limitations, this thesis explores lightweight and efficient pose-based approaches for human activity recognition using skeletal representations. This work presents a comprehensive study of pose-based HAR through two complementary approaches. A comparative analysis of classical machine learning models, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression (LR), and XGBoost, is conducted using statistical features extracted from MediaPipe-based skeletal keypoints. Experiments across benchmark datasets such as UCF101, PennAction, and HAA500 demonstrate that traditional machine learning models can achieve competitive performance with significantly lower computational requirements and a lightweight deep learning framework is proposed that leverages engineered skeletal features incorporating spatial joint coordinates, temporal motion dynamics, joint angles, and limb distances. Temporal modeling is performed using overlapping sliding windows, followed by dimensionality reduction using Principal Component Analysis (PCA) to retain 95% variance while significantly reducing feature dimensionality. A feed-forward neural network (FFNN) is employed for classification. The proposed method achieves high recognition performance, with up to 96% window-level accuracy and 98% video-level accuracy on UCF101, and comparable

results on PennAction.

The results demonstrate that well-designed skeletal feature representations combined with efficient machine learning and dimensionality reduction techniques can serve as a powerful alternative to computationally intensive deep learning models. This thesis highlights the potential of lightweight, pose-based HAR systems for real-time and resource-constrained environments, while maintaining strong recognition performance.

Keywords: Human Activity Recognition, Pose-based Activity Recognition, Sports Activity Recognition, MediaPipe, PCA

LIST OF PUBLICATIONS

Presented

INTERNATIONAL CONFERENCES

1. Aman Yadav, et al. "Comparative Analysis of Machine Learning Models for Sports Pose-Based Human Activity Recognition," *10th International Conference on Innovative Computing and Communication (ICICC-2026)*, Shaheed Sukhdev College of Business Studies, University of Delhi, India, in association with University of Valladolid, Spain. **(Accepted and presented; Springer LNNS Series, Scopus Indexed)**.

2. Aman Yadav, et al. "Lightweight Pose-Based Multi-Class Exercise Recognition Using PCA-Compressed Skeletal Features," *2nd Congress on Intelligent Machines and Algorithms (CIMA 2026)*, National Institute of Technology Puducherry, India. **(Accepted and Presented; Springer LNNS Series, Scopus Indexed)**.

Contents

ACKNOWLEDGEMENT i

CANDIDATE DECLARATION ii

CERTIFICATE iii

ABSTRACT iv

LIST OF PUBLICATIONS vi

LIST OF TABLES xii

LIST OF FIGURES xiii

LIST OF ABBREVIATIONS xiv

1 Introduction **1**

1.1 Background 2

1.2 Pipeline of Human Activity Recognition 2

1.2.1 Data Acquisition 3

1.2.2 Data Pre-processing 3

1.2.3 Feature Extraction 4

1.2.4 Activity Classification 4

1.3 Classification of Activity Recognition Approaches 4

1.3.1 Handcrafted Feature-Based Approaches 4

1.3.2 Deep Learning-Based Approaches 5

1.3.3 Hybrid Representation-Based Approaches 6

1.4 Activity Recognition Challenges 6
1.4.1 Occlusion 6
1.4.2 Viewpoint Variation 7
1.4.3 Dynamic Environments 7
1.4.4 Cluttered Background 7
1.4.5 Inter- and Intra-Class Variation 7
1.5 Applications 7
1.5.1 Video Surveillance 7
1.5.2 Healthcare and Rehabilitation 8
1.5.3 Sports Performance Analysis 8
1.5.4 Human–Computer Interaction 8
1.6 Motivation and Scope 9
1.7 Major Contributions of the Thesis 9
1.8 Organisation of the Thesis 10

2 Literature Review **12**
2.1 Human Pose Estimation 12
2.2 Skeleton-Based Human Activity Recognition 13
2.2.1 Handcrafted Skeletal Descriptors 13
2.2.2 Recurrent Neural Networks for Skeleton HAR 14
2.2.3 Graph Convolutional Networks 14
2.2.4 Transformer-Based Approaches 15
2.3 Classical Machine Learning for HAR 15
2.4 Dimensionality Reduction in HAR 16
2.5 Activity Recognition in Sports and Exercises 17
2.6 Lightweight and Deployable HAR Systems 17

2.7 Benchmark Datasets 18
2.8 Summary and Comparative Analysis 20
2.9 Research Gaps 20

3 Comparative Analysis of Machine Learning Models for Sports Pose-Based Human Activity Recognition 22

3.1 Introduction 22
3.2 Related Work 23
3.2.1 HAR in Sports 24
3.2.2 Machine Learning for Pose-Based HAR 24
3.3 Methodology 24
3.3.1 Preprocessing 25
3.3.2 Datasets and Selected Sports Classes 26
3.3.3 Pose Estimation and Feature Representation 27
3.3.4 Machine Learning Models 28
3.3.5 Evaluation Metrics 28
3.3.6 Implementation Details 29
3.4 Results Analysis 29
3.4.1 Results on UCF101 29
3.4.2 Results on PennAction 30
3.4.3 Results on HAA500 30
3.4.4 Overall Performance Analysis 30
3.5 Summary 32

4 Lightweight Pose-Based Multi-Class Exercise Recognition Using PCA-Compressed Skeletal Features 34

4.1 Introduction 34
4.2 Methodology (Pipeline and Skeleton Representation) 35
4.2.1 Skeleton Generation 35
4.2.2 Skeleton Preprocessing 36
4.2.3 Feature Extraction 37
4.2.4 Temporal Modeling 37
4.2.5 Dimensionality Reduction 38
4.2.6 Classification Model 38
4.3 Experimental Results 38
4.3.1 Datasets 38
4.3.2 Implementation Details and Evaluation Metrics 39
4.4 Result Analysis 39
4.4.1 Ablation Study 39
4.4.2 Window-Level Performance 40
4.4.3 Video-Level Performance 40
4.4.4 Discussion 41
4.5 Summary 41

5 Results and Analysis **43**
5.1 Overall Performance Comparison 43
5.2 Dataset-wise Analysis 43
5.3 Model Performance Comparison 44
5.4 Window-Level vs Video-Level Performance 44
5.5 Feature Effectiveness and Ablation Analysis 45
5.6 Computational Efficiency 45
5.7 Key Observations 45

5.8 Conclusion of Analysis 46
5.9 Statistical Reliability 46
5.10 Comparative Discussion 46

6 Conclusion, Future Scope and Societal Impact 47
6.1 Conclusion 47
6.2 Future Scope 47
6.3 Societal Impact 48

References 49

List of Publications and their Proofs 53

Plagiarism Report 63

Author's Biography 68

List of Tables

2.1 Summary of representative pose-based and skeleton-based HAR methods 19

3.1 Selected sports activity classes from benchmark datasets 27

3.2 Performance of ML models across all datasets (5-fold CV) 31

4.1 Ablation and Baseline Comparison (UCF101) 40

4.2 Window-Level Classification Performance on UCF101 and PennAction . 40

4.3 Video-Level Classification Performance on UCF101 and PennAction . . . 40

5.1 Overall Accuracy Comparison of Proposed Approaches 43

5.2 Performance of ML Models Across Datasets 44

5.3 Window-Level vs Video-Level Performance 45

5.4 Ablation Study on UCF101 (PCA-Based Model) 45

List of Figures

1.1 Workflow diagram for a Human Activity Recognition system 3

1.2 Key challenges and problems in activity recognition 6

1.3 Applications of Human Activity Recognition 8

3.1 Overview of the proposed pose-based HAR pipeline including MediaPipe pose extraction, feature engineering, and ML-based classification 25

3.2 MediaPipe 33-Keypoint Skeleton Representation 26

3.3 Performance comparison on UCF101 dataset 29

3.4 Performance comparison on PennAction dataset 30

3.5 Performance comparison on HAA500 dataset 31

3.6 Performance comparison of the best ML models across the datasets 32

4.1 Selected 13-joint skeleton configuration 36

4.2 Overview of the proposed pose-based exercise recognition pipeline . . . 37

4.3 (Left) PCA dimensionality reduction from 2040 to 120/91. (Right) Accuracy vs. feature dimensions around 120 dimensions 41

List of Abbreviations

AI Artificial Intelligence

HAR Human Activity Recognition

ML Machine Learning

DL Deep Learning

CNN Convolutional Neural Network

RNN Recurrent Neural Network

LSTM Long Short-Term Memory

FFNN Feed-forward Neural Network

SVM Support Vector Machine

KNN K-Nearest Neighbors

RF Random Forest

LR Logistic Regression

PCA Principal Component Analysis

RGB Red Green Blue

FPS Frames Per Second

GNN Graph Neural Network

GCN Graph Convolutional Network

GAT Graph Attention Network

ST-GCN Spatial Temporal Graph Convolutional Network

CHAPTER 1
INTRODUCTION

Activity Recognition, a sub-domain of vision-related applications, is the ability to identify and recognise the actions or goals of an agent. The agent can be any object or entity that performs an action with end goals, operating either individually or as part of a group. One such agent is the human being, and recognising the activities of humans is referred to as *Human Activity Recognition* (HAR). The goal of activity recognition is to automatically analyse ongoing activities from an unknown video or image sequence. In a simple case where a video is segmented to contain only one execution of a human activity, the objective of the system is to correctly classify the video into its activity category [19].

In recent years, HAR has gained significant momentum in the computer vision community, particularly for applications in intelligent fitness monitoring, sports performance analysis, clinical rehabilitation, and human–computer interaction [20, 42]. While state-of-the-art deep learning models such as TimeSformer [7] and Video Swin Transformer [32] demonstrate impressive accuracy on large-scale video benchmarks, they operate directly on raw RGB pixel data, making them computationally expensive, sensitive to illumination and background variation, and unsuitable for deployment on resource-constrained edge devices.

A potential approach is that of *pose-based* HAR, where human activity can be described in terms of key joints in the skeleton form without using pixel information [30]. By disregarding visual details of the subjects, pose-based approaches inherently become more robust to irrelevant factors like occlusions, viewpoints, and changes in illumination. In addition, the use of visual details for activity classification does not compromise the privacy of the subjects. Modern pose estimation algorithms such as MediaPipe BlazePose [6], OpenPose [9], and AlphaPose [16] provide efficient ways to extract key skeletons in real-time conditions.

This thesis presents two complementary contributions in the domain of pose-based HAR for sports and exercise. First, a systematic comparative study of five classical machine learning classifiers Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), Logistic Regression (LR), and XGBoost is conducted on three benchmark datasets (UCF101, PennAction, HAA500) using a unified 396-dimensional statistical pose descriptor. Second, a lightweight multi-class exercise recognition framework is proposed that combines hybrid engineered skeletal features, PCA-based temporal compression, and a shallow feed-forward neural network (FFNN) evaluated under a strict video-level split to prevent data leakage. Together, these contributions demonstrate that

well-designed, compact skeletal representations can deliver strong recognition performance without the complexity of graph-based or transformer-based architectures.

1.1 Background

Activity recognition from images or videos is a challenging field with real-world impact, drawing increasing attention from the research community. The goal of HAR is to automatically analyse ongoing activities from an unknown video that is, a sequence of image frames. In more general cases, continuous recognition of human activities requires detecting starting and ending times of all occurring activities within an input video stream. Activity recognition finds utility in surveillance systems, patient monitoring, sports analytics, and human–computer interaction.

Due to significant variation in high-dimensional video data, cluttered backgrounds, partial occlusion, and varying motion speed, explicit action detection and recognition remain a major challenge [63]. Efficient solutions to these problems enable useful applications in monitoring, visual surveillance, human–robot cooperation, and medical monitoring systems. The intensity of these challenges varies depending on the category of activity under consideration.

Depending on their complexity and duration [20], human activities are conceptually categorised into four levels: *gestures*, *actions*, *interactions*, and *group activities*.

Gesture: A gesture is a basic movement of human body parts that carries semantic meaning. Head shaking, hand waving, and facial expressions are common examples. Gestures are typically brief and represent the lowest level of complexity.

Action: An action is an activity performed by a single person, composed of multiple gestures (atomic actions). Walking, running, jogging, squatting, and push-ups are examples. In the context of this thesis, structured exercise actions jumping jacks, pull-ups, push-ups, squats, lunges, and bench press constitute the primary recognition targets.

Interaction: An interaction is an activity involving two actors, at least one of whom is human. Human–human interactions include handshaking and fighting human–object interactions include a person using a computer or lifting a weight.

Group Activity: The most complex category, involving more than two humans and potentially multiple objects. A team performing a coordinated warm-up drill or two teams competing in a sport are examples of group activities.

1.2 Pipeline of Human Activity Recognition

Typically, an HAR system consists of five stages as illustrated in Figure 1.1: (1) input data capture, (2) data pre-processing, (3) model building and training, (4) activity recognition, and (5) result output or user interface. Each stage admits multiple implementation strategies, and the choice of application domain, data acquisition device, and AI algorithm

makes the design space both rich and challenging.

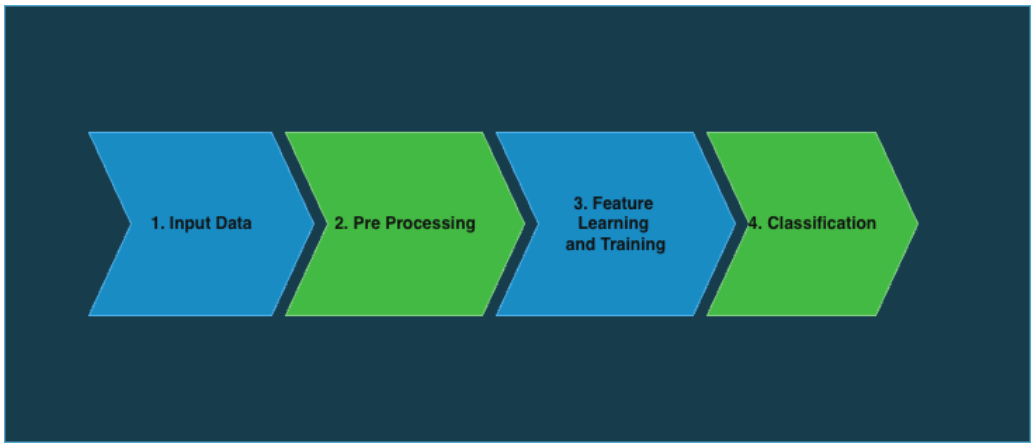

Figure 1.1: Workflow diagram for a Human Activity Recognition system

1.2.1 Data Acquisition

The input to an HAR system can originate from RGB cameras, depth sensors (RGB-D), inertial measurement units (IMUs), or combinations thereof. In this thesis, standard monocular RGB video is used as input across all three benchmark datasets (UCF101, PennAction, HAA500). MediaPipe BlazePose [6] is applied frame-by-frame to extract 33 body landmark keypoints per frame, producing a compact skeletal representation that serves as the sole input to all downstream models.

1.2.2 Data Pre-processing

Raw pose keypoints are subject to noise from detection failures, camera placement differences, and inter-subject size variation. Two normalisation steps ensure spatial consistency. *Translation normalisation* relocates all joint coordinates so that the neck joint (midpoint of the shoulders) serves as the origin, eliminating camera-position-induced spatial bias. *Scale normalisation* divides all coordinates by the average torso length (neck-to-hip distance), removing the effect of subject-to-camera distance and subject scale variation. Frames with unreliable pose detections (missing keypoints) are discarded this affected fewer than 5% of clips across all datasets and did not materially affect results.

1.2.3 Feature Extraction

Two complementary feature representations are employed in this thesis.

For the *comparative ML study* (Chapter 3), temporal statistical descriptors mean, standard deviation, minimum, and maximum are computed over all frames for each of the 33 landmark coordinates (x, y, z), yielding a fixed-length 396-dimensional vector per video clip ($33 \times 3 \times 4 = 396$). This representation captures the temporal envelope of each joint trajectory without requiring explicit sequence modelling.

For the *lightweight FFNN framework* (Chapter 4), a hybrid frame-level feature vector of 68 dimensions is constructed by combining: (i) 26 normalised spatial joint coordinates, (ii) 26 frame-to-frame joint velocity features, (iii) 8 biomechanical joint angle features, and (iv) 8 limb distance features. Overlapping sliding windows (30 frames, stride 15) then flatten these into 2040-dimensional temporal vectors, which are compressed to 120 (UCF101) or 91 (PennAction) dimensions via Principal Component Analysis retaining 95% of the variance.

1.2.4 Activity Classification

In the comparative study, five classical ML classifiers (SVM, RF, KNN, LR, XGBoost) are evaluated using 5-fold stratified cross-validation. In the lightweight framework, a shallow three-hidden-layer FFNN (256–128–64 neurons, ReLU activations, dropout 0.3) is trained with the Adam optimiser under a strict video-level 80/20 split. Majority voting over window-level predictions yields final video-level decisions [25].

1.3 Classification of Activity Recognition Approaches

Vision-based HAR approaches are broadly classified into three categories: (i) handcrafted feature-based, (ii) deep learning representation-based, and (iii) hybrid representation-based approaches.

1.3.1 Handcrafted Feature-Based Approaches

Handcrafted feature-based methods represent the conventional paradigm for action recognition. These methods analyse video sequences or frames to extract descriptive features and construct a representation vector that is passed to a classifier. They rely on domain knowledge and human ingenuity, and are categorised as space-time based or appearance based.

Space-Time Based Approaches

Space-time approaches characterise actions using spatiotemporal features extracted at Space-Time Interest Points (STIPs). Both dense and sparse STIP detectors are employed dense detectors cover the entire video content, while sparse detectors use subsets to identify interest points. Descriptors such as HOG, HOF, and MBH capture local or global spatial and motion information from these points [63].

Appearance-Based Approaches

Appearance-based methods exploit shape, motion, or both for action representation. Shape-based approaches use contour and silhouette features motion-based approaches rely on optical flow or its variants. Examples include Shape Histogram, Body Skeleton representations, Motion History Volume, and Optical Flow [63]. The pose-based feature representations used in this thesis can be viewed as a principled evolution of skeletal appearance representations, replacing handcrafted pixel-level descriptors with structured, body-centric keypoint features.

1.3.2 Deep Learning-Based Approaches

Deep learning has substantially transformed HAR by enabling end-to-end feature learning directly from raw data, removing the need for handcrafted descriptors. These models automatically discover hierarchical representations through multiple non-linear layers, and are categorised as generative, discriminative, or hybrid.

Generative Model-Based Approaches

Generative models learn the distribution of activity data in an unsupervised or semi-supervised manner. Deep Belief Networks (DBNs), Autoencoders, and Generative Adversarial Networks (GANs) are commonly used to learn compact, class-conditional latent representations of activity sequences.

Discriminative Model-Based Approaches

Discriminative models use supervised deep hierarchies to map raw inputs to activity labels. CNNs learn spatial features from individual frames, while RNNs and LSTM networks model temporal dependencies across frames [30, 42]. Graph Convolutional Networks such as ST-GCN [56] and CTR-GCN [12] extend convolution to the skeletal graph topology, jointly learning spatial and temporal dependencies. The ablation study in Chapter 4 shows that a lightweight FFNN with PCA-compressed temporal features (96.0% accuracy on UCF101) outperforms an LSTM baseline (93.5%), demonstrating that recurrent complexity is not always necessary for well-structured exercise classes.

1.3.3 Hybrid Representation-Based Approaches

Hybrid methods involve manual and deep learning-based representations, each having their own advantage that is leveraged by the combination. For example, the combination of HOG or optical flow descriptors along with CNN based features has proven to provide better results for tasks under partial occlusions and low resolutions. The lightweight method suggested in Chapter 4 can be classified as an optimal hybrid in which biomechanical features are used with PCA dimensionality reduction and a shallow neural classifier.

1.4 Activity Recognition Challenges

Even with advancements made, HAR in natural scenarios is still an open problem. Some of these are depicted in Figure 1.2.

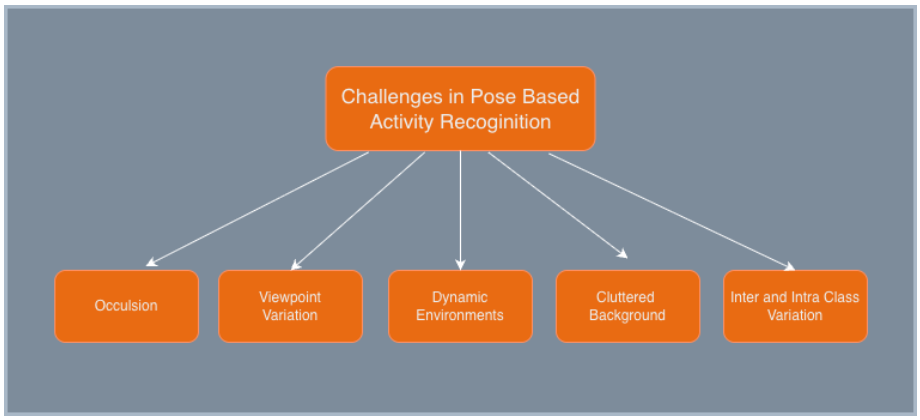

Figure 1.2: Key challenges and problems in activity recognition

1.4.1 Occlusion

It must be possible to see an activity that needs to be recognized through a video sequence. Practically, depending on whether the scene is a surveillance or a gym, it is likely that parts of the body would be occluded either by the person itself or by other factors. This makes it impossible to detect joints which leads to poor pose estimation results and hence classification. In the proposed thesis, such frames are filtered out during the preprocessing stage.

1.4.2 Viewpoint Variation

A given exercise might have different appearances depending on the perspective in which the scene is shot. From different perspectives such as frontal, lateral, or overhead, a given exercise would have varying joint coordinates. Translation normalization partially tackles this problem.

1.4.3 Dynamic Environments

Varying lighting conditions, moving backgrounds, and non-stationary cameras introduce noise into both the pose estimation step and the resulting skeletal features. Pose-based methods are inherently more robust to environmental dynamics than RGB-based methods, since they abstract away pixel-level appearance variation.

1.4.4 Cluttered Background

Cluttered backgrounds introduce ambiguous motion information that can confuse optical flow-based descriptors and appearance-based detectors. By operating on skeleton key-points rather than raw frames, the proposed systems in this thesis are largely insensitive to background clutter a key advantage over pixel-level approaches.

1.4.5 Inter- and Intra-Class Variation

Structurally similar exercises such as squats and lunges, or push-ups and bench press share overlapping joint motion patterns, making them challenging to discriminate.

1.5 Applications

1.5.1 Video Surveillance

HAR enables intelligent surveillance systems capable of automatically detecting anomalous or suspicious behaviours in public spaces, reducing the burden on human operators and enabling faster incident response.

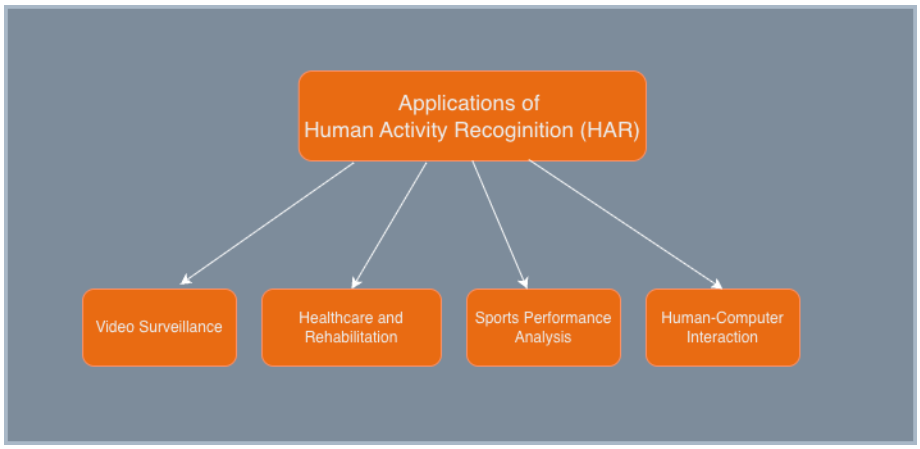

Figure 1.3: Applications of Human Activity Recognition

1.5.2 Healthcare and Rehabilitation

In clinical and home-based rehabilitation settings, automated exercise recognition can provide real-time feedback on form, count repetitions, and monitor patient progress over time. Pose-based systems are particularly attractive here due to their privacy-preserving nature no identifiable RGB imagery needs to be stored or transmitted.

1.5.3 Sports Performance Analysis

Automated recognition and quantitative analysis of athletic movements enables objective performance monitoring, biomechanical assessment, injury risk detection, and personalised coaching feedback [8, 10]. The exercise activities targeted in this thesis squats, push-ups, pull-ups, lunges, jumping jacks, and bench press are representative of this domain.

1.5.4 Human–Computer Interaction

Gesture and action recognition underpins touchless interfaces, augmented reality, and robotic control via natural body movement. Lightweight pose-based pipelines are well-suited for real-time interaction on edge devices such as smart glasses or embedded microcontrollers.

1.6 Motivation and Scope

There are many compelling applications of activity recognition systems. HAR plays a key role in healthcare, sports analytics, surveillance, human–computer interaction, and smart environments. One of the central challenges is that the same activity can be performed differently by different individuals, or even by the same individual on different occasions. Finding a discriminative representation that separates distinct activities while grouping together varied executions of the same action remains a fundamental research problem [63].

The key motivations for this thesis are:

- The rapid expansion of fitness monitoring devices, smart gymnasium infrastructure, and telehealth platforms creates an urgent demand for accurate, real-time exercise recognition systems deployable on resource-constrained hardware. RGB-based deep models are impractical in these settings due to their computational demands and privacy concerns hence deploying real-time multi-spectral sensor fusion on edge microcontrollers and mobile devices remains a critical priority for modern, lightweight skeleton-based frameworks[27, 60].

- Despite the growing body of work on pose-based HAR, few studies provide systematic, controlled comparisons of classical machine learning classifiers across multiple sports-oriented datasets under a consistent feature representation and evaluation protocol.

- In the current existing state, many lightweight HAR techniques have been proposed, but most are not supported by biomechanically motivated feature engineering, leakage-free and robust evaluation, or statistical rigor. These limitations are addressed in this thesis.

- Activity recognition technology has the potential to make a significant societal impact by enabling applications such as assisted living for the elderly, early detection of abnormal behaviours, and personalised home-based rehabilitation.

The scope of this thesis is restricted to single-person, structured exercise activities under semi-controlled recording conditions, evaluated on publicly available benchmark datasets. Extensions to multi-person scenarios, in-the-wild settings, and 3D pose representations are identified as future research directions.

1.7 Major Contributions of the Thesis

The major contributions of this thesis are:

1. **Comparative ML Study for Sports Pose-Based HAR.** A comprehensive comparative evaluation of five classical ML classifiers using a unified 396-dimensional statistical pose descriptor on UCF101, PennAction, and HAA500 under 5-fold stratified cross-validation. SVM (RBF) achieves the best overall performance (98.1% on UCF101, 93.0% on HAA500) XGBoost and LR are jointly best on PennAction (95.4%).

2. **Hybrid Skeletal Feature Representation.** A 68-dimensional frame-level feature vector integrating spatial joint coordinates, temporal motion, biomechanical joint angles, and limb distances that captures complementary kinematic information beyond raw coordinate data alone.

3. **PCA-Compressed Temporal Modelling.** A sliding-window scheme (30 frames, stride 15) combined with PCA compression ($2040 \rightarrow 120/91$ dimensions for two different datasets, 95% variance retained), achieving over 94% dimensionality reduction without measurable loss of discriminative information.

4. **Lightweight FFNN with Leakage-Free Evaluation.** A shallow three-hidden-layer FFNN trained under strict video-level splitting achieves 96.0% window-level and 98.25% video-level accuracy on UCF101, and 95.5% and 95.0% on PennAction, demonstrating that complex graph-based architectures are unnecessary for structured exercise recognition.

5. **Statistical Reliability Analysis.** Multi-seed evaluation yields window-level accuracy of $96.77\% \pm 0.22\%$ (95% CI: [96.50%, 97.03%]) and video-level accuracy of $98.09\% \pm 0.85\%$ (95% CI: [97.03%, 99.14%]). Bootstrap resampling further confirms the robustness of the reported results.

1.8 Organisation of the Thesis

The remainder of the thesis is organised as follows.

Chapter 2 provides a comprehensive literature review of human pose estimation, skeleton-based HAR, classical ML for action recognition, and deep learning architectures for pose-based activity classification, along with a review of relevant benchmark datasets.

Chapter 3 presents the comparative analysis of ML models for sports pose-based HAR, detailing the pipeline, feature representation, ML models, datasets, and evaluation methodology, followed by experimental results on UCF101, PennAction, and HAA500.

Chapter 4 presents the lightweight pose-based multi-class exercise recognition framework, covering hybrid skeletal feature engineering, sliding-window temporal modelling, PCA compression, and the lightweight FFNN classifier, including ablation studies, window-level and video-level performance, and statistical reliability analysis.

Chapter 5 provides the results and analysis of the performance comparison, performance analysis of individual datasets, computational efficiency, ablation testing, and statistical reliability tests of the proposed methods.

Chapter 6 concludes the thesis by summarising findings, discussing social impacts and limitations, and outlining future research directions including 3D pose representations, large-scale dataset evaluation, and lightweight temporal attention mechanisms.

CHAPTER 2

LITERATURE REVIEW

The recognition of human activities from visual data has been an active research area for several decades, evolving from rule-based and handcrafted-feature methods to sophisticated deep learning architectures. This chapter presents a structured review of prior work organised around the key topics that underpin the contributions of this thesis: human pose estimation, skeleton-based HAR, classical machine learning for action recognition, deep learning architectures for HAR, sports and exercise recognition, and lightweight deployable HAR systems. Within each topic, works are reviewed in broadly chronological order to trace the evolution of ideas. The chapter concludes with a comparative summary table and a discussion of the gaps identified in the literature.

2.1 Human Pose Estimation

Human pose estimation is the foundational prerequisite of any skeleton-based HAR system. The objective is to localise anatomical body joints shoulders, elbows, wrists, hips, knees, and ankles in image or video data and represent their configuration as a structured skeleton.

Early approaches formulated pose estimation as inference in pictorial structure models. Fischler and Elschlager [18] introduced the deformable part model concept, representing the body as a collection of parts connected by spring-like constraints. Lateer on Felzenszwalb and Huttenlocher [17] demonstrated efficient inference in these models utilising dynamic programming. While these methods were elegant but struggled with complex poses and multiple persons due to their reliance on hand-designed part detectors.

The introduction of deep convolutional networks transformed pose estimation. Toshev and Szegedy [48] proposed DeepPose, which regressed joint coordinates directly from image patches using a cascade of CNN regressors, which significantly outperformed the prior methods on benchmark datasets. Stacked hourglass network was introduced by Newell et al. [35], which applied a symmetric encoder-decoder architecture with intermediate supervision to produce joint heatmaps at multiple scales, which became one of the most widely adopted baselines in the field.

Multi-person pose estimation was advanced by Cao et al. [9], whose OpenPose framework introduced Part Affinity Fields (PAFs) to associate detected joints with individual subjects in a bottom-up fashion, which enabled real-time multi-person estimation

without the requirement of any person detectors. In the 2019 Sun et al. [46] proposed HRNet, which maintains a high-resolution representations throughout the network rather than recovering resolution from a low-resolution bottleneck, achieving state-of-the-art single-person pose estimation accuracy on COCO.

For on-device and real-time applications, Bazarevsky et al. [6] developed MediaPipe BlazePose, a two-stage pipeline that first detects the subject with a lightweight detector and then regresses 33 body landmarks using a single forward pass. BlazePose achieves real-time performance on mobile and embedded hardware, making it the pose extraction backbone of choice for the systems proposed in this thesis. Each keypoint is provided with normalised (x, y) coordinates and a relative depth offset z, which, although not a true 3D metric depth, encodes useful relative limb depth information estimated from monocular RGB input.

2.2 Skeleton-Based Human Activity Recognition

Skeleton-based HAR represents human motion as a temporal sequence of body joint coordinates, abstracting away appearance information and enabling recognition that is robust to changes in clothing, illumination, and background. The field has progressed from handcrafted skeletal descriptors to graph convolutional and transformer-based architectures.

2.2.1 Handcrafted Skeletal Descriptors

Early skeleton-based methods manually engineered features from joint trajectories to capture discriminative motion patterns. Ofli et al. [36] proposed the Sequence of the Most Informative Joints (SMIJ), selecting the most discriminative joints for each activity and computing statistics over their trajectories. Vemulapalli et al. [49] represented skeleton sequences as curves in a Lie group of rigid-body transformations, capturing both joint angles and limb lengths in a geometrically principled way, and achieved strong results on action recognition benchmarks using dynamic time warping and Fourier temporal pyramid matching.

Zanfir et al. [58] introduced the Moving Pose descriptor, which combines joint position, velocity, and acceleration into a per-frame descriptor and uses a modified nearest-neighbour classifier for recognition. The importance of temporal dynamics particularly velocity and acceleration features noted in these early works is reflected in the hybrid feature representation proposed in this thesis, where frame-to-frame joint velocity features form a key component of the 68-dimensional frame-level descriptor.

2.2.2 Recurrent Neural Networks for Skeleton HAR

The sequential nature of skeleton data makes recurrent architectures a natural fit. Shahroudy et al. [41] introduced the NTU RGB+D dataset and proposed a part-aware LSTM that processes different body parts through separate memory cells, capturing part-level temporal dynamics independently. Liu et al. [30] extended this with a spatio-temporal LSTM architecture incorporating trust gates that adaptively weight the contribution of noisy joints, achieving strong performance on both NTU RGB+D and smaller benchmarks.

Sharma and Singh [42] proposed ConvST-LSTM-Net, which combines convolutional spatial feature extraction with spatiotemporal LSTM layers to jointly model spatial and temporal dependencies in skeleton sequences. Singh and Sharma [44] further introduced STAD-ConvBi-LSTM, a spatio-temporal attention mechanism within a deep convolutional bidirectional LSTM framework for abnormal activity detection, demonstrating strong generalisation across diverse action categories.

2.2.3 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) are inherently able to model the skeletal topology in the form of a graph, where joints serve as nodes, and physical connectivity serves as edges, hence facilitating joint modelling of spatial and temporal aspects.

ST-GCN is the first trainable GCN proposed by Yan et al. [56] for skeleton-based HAR and defines spatial graph convolutions over the joint adjacency matrix as well as temporal convolutions across the time axis. Even with using the fixed and physically-defined graph topology, ST-GCN proved its efficiency significantly surpassing previous handcrafted techniques on NTU RGB+D and Kinetics-Skeleton datasets. Shi et al. [43] improved the ST-GCN architecture introducing a two-stream adaptive graph convolutional network (AAGCN) that learns data-driven graph structures along with the physically-defined one, hence capturing the long-range dependencies between joints not present in the fixed graph structure. Chen et al. [12] proposed an improvement to this idea by introducing CTR-GCN and refining the connectivity structure in a channel-wise manner, therefore allowing each feature channel to model different joint relations, resulting in state-of-the-art performance on NTU RGB+D 60 and 120.

Efforts toward efficient GCN models involve efficient temporal convolutions [3], depthwise separable attention to reduce the quadratic complexity [26]. Furthermore, the topological issues are tackled by differentiating central and edge nodes to prevent over-smoothing [51], refining the adjacency matrix to prevent bone connectivity degradation [65], and applying heuristic initializations of the model [53]. Finally, improvements can be made on the skeletal topology representation via contrastive information bottlenecks [13] as well as from the prototypical perspective [29].

Although efficient, GCN-based models come with huge parameter costs and complexity which make them inapplicable in constrained deployments, being a disadvantage

specifically tackled by the lightweight framework of this thesis.

2.2.4 Transformer-Based Approaches

Transformer architectures have recently been adapted for skeleton-based HAR to model long-range temporal dependencies through self-attention. Plizzari et al. [37] proposed spatial and temporal transformer networks that apply self-attention within each frame across joints (spatial) and across frames for each joint (temporal), capturing both intra-frame joint relationships and inter-frame motion patterns.

For video-level HAR from RGB frames, Bertasius et al. [7] proposed TimeSformer, applying divided space-time attention that separately attends over spatial patches and temporal positions. Liu et al. [32] proposed the Video Swin Transformer, adapting shifted window attention from the image domain to spatio-temporal video volumes. While these models achieve top accuracy on large-scale benchmarks, they require high GPU memory and are unsuitable for on-device deployment.

2.3 Classical Machine Learning for HAR

Classical machine learning classifiers occupied the dominant position in HAR before the deep learning era and have experienced a resurgence of interest for lightweight, deployable systems. Their computational efficiency, interpretability, and strong performance on compact feature representations make them well-suited to pose-based exercise recognition.

Support Vector Machines (SVMs) have been extensively applied to HAR because of their ability to find maximum-margin decision boundaries in high-dimensional feature spaces. Schuldt et al. [40] demonstrated the effectiveness of the SVMs with spatial-temporal video features for action recognition, establishing SVMs as a strong baseline in the field and subsequent work confirmed that SVMs with RBF kernels consistently outperform other classical classifiers when paired with skeletal or pose-based features [64].

Several works have applied Random Forests to body-sensor-based and pose-based HAR, and reporting competitive accuracy with lower training time compared to SVMs in certain settings [10]. Ensemble boosting methods such as the AdaBoost and more recently, XGBoost [11] have further advanced ensemble-based classification by iteratively focusing on misclassified samples and applying gradient-based loss minimisation.

K-Nearest Neighbours (KNN) provides a non-parametric baseline that requires no training phase but it suffers from the curse of dimensionality and the high inference latency for large feature spaces.

Logistic Regression with ℓ_2 regularisation serves as a strong linear baseline and is particularly informative for assessing the feature separability that is if LR achieves high

accuracy, the feature distributions are largely linearly separable, which simplifies the classification problem.

Zhao and Zhang [64] conducted a comprehensive comparative study of SVM, RF, KNN, and LR for pose-based HAR, concluding that SVM (RBF) is the most reliable classifier across datasets and feature dimensionalities. Cai et al. [8] applied SVM and RF to the MediaPipe-extracted skeletal features for fitness activity recognition, confirming that pose-based statistical descriptors combined with classical classifiers are competitive with the deep learning approaches for structured exercise classes. These findings motivate the comparative study presented in Chapter 3 of this thesis.

2.4 Dimensionality Reduction in HAR

The high dimensionality of raw skeletal feature vectors particularly those derived from sliding-window temporal modelling motivates the use of dimensionality reduction to remove redundancy and improve classifier generalisation.

Principal Component Analysis (PCA) is the most widely adopted linear dimensionality reduction technique in machine learning. By projecting data onto the directions of maximum variance, PCA removes correlated and low-variance dimensions while preserving the dominant structure of the data. In the context of HAR, PCA has been applied to joint trajectory features [36], optical flow descriptors, and temporal feature aggregates to reduce computational overhead without degrading recognition accuracy.

In this thesis lightweight framework, the use of PCA reduces the sliding-window temporal feature vector size from 2040-dimensions to 120 dimensions (UCF101) and 91 dimensions (PennAction), which accounts for 95% of the variance while achieving a compression ratio of 17:1. From the discussion in Chapter 4, the use of dimensionality reduction in this case helps in lowering inference costs and enhancing generalisation by reducing noisy dimensions, thus resulting in a model that outperforms the model trained without it.

PCA can still prove useful beyond the scope of transforming spatial coordinates. For example, some federated learning models such as CF-STGAT [57] employ PCA to transform attention descriptors and stabilise client drifts. Moreover, studies concerning manual material handling apply PCA-transformed inputs for highly accurate motion sequence detection [24]. PCA can also be used in other fields to compress word embeddings while preserving action-related semantics [1].

t-SNE and UMAP are examples of non-linear dimensionality reduction methods that could be employed in visualisation but cannot be applied to inference data since their solutions depend heavily on context. Autoencoders have been applied in HAR for dimensionality reduction but require more resources due to the need for an encoder, which does not necessarily generalise across datasets. PCA provides a closed form solution and deterministic results, thus making it more preferable for the task at hand.

2.5 Activity Recognition in Sports and Exercises

Activity recognition in the domain of sports and exercises has received increasing attention due to its potential applications in performance monitoring, coaching, and rehabilitation assessment.

Initially, activity recognition involved the use of body-worn sensors. For example, Bao and Intille [5] were able to accurately recognize exercise and daily activities through a combination of accelerometry features and nearest neighbor and decision tree classifiers. The research into various other sensing modalities including gyroscopes, insole pressure sensors, and electromyography showed the advantages of sensor fusion and multimodal sensing even though the complexity of the required hardware increases.

The vision-based activity recognition has evolved from traditional RGB frame classification towards pose-based solutions as efficient pose estimation became available. For example, Chen et al. [10] used pose keypoints extracted by MediaPipe algorithm to accurately classify push-ups, squats, and sit-ups using simple classifiers. Similarly, the SVM and Random Forest algorithms have been successfully employed by Cai et al. [8] for recognizing gym exercises based on pose data, achieving over 90% accuracy in classification.

Several recent frameworks have been developed specifically for various sports. For example, SportPoseNet combines MediaPipe pose detection with a refined ResNet-200 architecture to classify sports actions with very high accuracy [34]. Hybrid CNN and RNN architectures combined with the key body landmarks also proved to be effective for more complex activity classification [47]. Architectures like ASTM-Net utilize multi-branch graph convolution and dynamic modeling of Activity-Aware Adjacency Graphs (3A-Graphs) to accommodate different speeds and durations of various sports actions [52].

Some works focused specifically on exercise repetition counting and quality assessment rather than classification. The pose-based activity representation is especially advantageous in this setting since joint angles and velocities provide a natural representation of movement quality. For example, the knee flexion angle in a squat or the elbow extension angle in a push-up represent the exercise form directly. The proposed hybrid representation explicitly includes the joint angles and limb distances among the body landmarks.

Benchmark datasets for sports and exercise HAR include UCF101 [45], which contains 101 action classes including sports activities PennAction [61], which provides frame-level joint annotations for 15 classes and HAA500 [14], a human-centric atomic action dataset with 500 fine-grained activity categories. The comparative study in Chapter 3 evaluates all three datasets using five sports-oriented activity classes selected from each.

2.6 Lightweight and Deployable HAR Systems

The deployment of HAR systems on edge devices smartphones, smart glasses, gymnasium sensors, and rehabilitation wearables imposes strict constraints on model size, inference latency, and memory footprint. This has motivated a growing body of work on lightweight

HAR architectures.

MobileNet [21] introduced depthwise separable convolutions to reduce the parameter count and FLOPs of image classification networks by an order of magnitude. Subsequent lightweight video architectures such as MobileNetV3 and EfficientDet adapted similar principles for temporal modelling. However, even lightweight video CNNs operate on raw RGB frames and impose higher computational demands than skeleton-based methods operating on compact keypoint sequences.

For skeleton-based HAR, Zhang et al. [63] identified that graph-based and transformer-based architectures, despite their high accuracy, are overparameterised for structured exercise activities with limited class sets and distinctive biomechanical patterns. Lighter alternatives including simple MLPs on flattened skeleton sequences and CNN-based methods on skeleton image representations have been explored, with competitive results on constrained activity sets.

The work of Plizzari et al. [37] proposed efficient transformer variants for skeleton HAR with reduced attention complexity, but the overall model still requires GPU inference. In contrast, the lightweight FFNN proposed in this thesis requires only CPU inference and achieves 96.0% window-level and 98.25% video-level accuracy on UCF101, operating on 120-dimensional PCA-compressed features derived from a 13-joint skeletal representation. This positions the proposed framework as one of the most computationally efficient competitive methods for exercise-specific HAR in the literature.

2.7 Benchmark Datasets

A brief review of the benchmark datasets used in this thesis is presented below.

UCF101 [45] contains 13,320 video clips across 101 human action classes collected from YouTube. The dataset covers a wide range of activities including sports, daily life, and exercise movements. For this thesis, five exercise classes are selected: Body Weight Squats, Jumping Jacks, Lunges, Pull-Ups, and Push-Ups, comprising 564 videos. UCF101 is challenging due to wide variation in camera angles, background clutter, and performer styles.

PennAction [61] contains 2,326 video clips of 15 sport and exercise categories with per-frame joint annotations. The dataset includes fine-grained temporal annotations at the sub-action (acteme) level. For this thesis, five classes are selected: Jumping Jacks, Pull-Up, Push-Up, Squat, and Bench Press, comprising 853 samples. PennAction poses inter-class confusion challenges due to the structural similarity between upper-body exercises (push-up vs. bench press) and between lower-body exercises.

HAA500 [14] is a recently introduced human-centric atomic action dataset containing 500 fine-grained activity categories, each with 20 curated high-quality video clips. For this thesis, five classes are selected: Jumping Jack, Pull-Ups, Push-Ups, Sit-Up, and Lunge, comprising 100 videos. HAA500 is the most challenging of the three datasets due to its small per-class sample count and fine-grained activity distinctions.

Table 2.1: Summary of representative pose-based and skeleton-based HAR methods

Method	**Classifier / Architecture**	**Dataset(s)**	**Accuracy (%)**	**Year**
Vemulapalli et al. [49]	SVM on Lie group features	MSR-Action3D, G3D	88.2 / 92.0	2014
Shahroudy et al. [41]	Part-aware LSTM	NTU RGB+D	62.9	2016
Yan et al. [56]	ST-GCN	NTU RGB+D, Kinetics	81.5 / 30.7	2018
Liu et al. [30]	Spatio-temporal LSTM	NTU RGB+D	69.2	2020
Shi et al. [43]	Two-stream AAGCN	NTU RGB+D 60	88.0	2019
Chen et al. [12]	CTR-GCN	NTU RGB+D 60/120	92.4 / 88.9	2021
Plizzari et al. [37]	Spatial-temporal Transformer	NTU RGB+D 60	89.9	2022
Sharma & Singh [42]	ConvST-LSTM-Net	NTU RGB+D	90.1	2023
Cai et al. [8]	SVM / RF on MediaPipe features	Custom fitness	>90.0	2022
Zhao & Zhang [64]	SVM, RF, KNN, LR	Multiple pose datasets	Up to 94.0	2022
Yadav & Sharma (This thesis, Ch. 3)	SVM, RF, KNN, LR, XGBoost	UCF101, PennAction, HAA500	**98.1 / 95.4 / 93.0**	2026
Yadav & Sharma (This thesis, Ch. 4)	FFNN + PCA	UCF101, PennAction	**98.25 / 95.0**	2026

2.8 Summary and Comparative Analysis

Table 2.1 summarises representative works in pose-based and skeleton-based HAR, including their methodology, datasets, and reported accuracy, to contextualise the contributions of this thesis.

The review reveals several consistent trends. First, deep learning architectures particularly GCNs and transformers achieve the highest accuracy on large-scale general action benchmarks such as NTU RGB+D and Kinetics, but their computational demands and large model sizes make them unsuitable for deployment on edge devices. Second, classical machine learning classifiers combined with well-engineered pose descriptors deliver competitive accuracy on structured activity sets particularly exercise and sports classes at a fraction of the computational cost. Third, rigorous evaluation protocols are essential: many reported results in the literature are inflated by data leakage arising from window-level splits that allow frames from the same video to appear in both training and test sets.

The contributions of this thesis directly address these observations by demonstrating that: (i) a unified statistical pose descriptor combined with five classical classifiers achieves up to 98.1% accuracy on sports-specific benchmarks and (ii) a hybrid engineered feature representation with PCA compression and a shallow FFNN achieves 98.25% video-level accuracy on UCF101 under a strict leakage-free evaluation protocol results that are competitive with GCN-based methods at substantially lower complexity.

2.9 Research Gaps

Drawing on the review above, the following specific gaps in the existing literature directly motivate the contributions of this thesis:

1. While several comparative studies of classical ML classifiers for HAR exist, none provides a controlled, multi-dataset comparison specifically targeting sports and exercise activities under a unified pose descriptor and consistent 5-fold stratified cross-validation protocol.

2. The integration of biomechanically motivated features joint angles, limb distances, and temporal motion within a PCA-compressed sliding-window representation for lightweight exercise recognition has not been previously explored.

3. Most sliding window HAR studies fail to ensure video-level separation between training and test data, leading to data leakage. The use of video-level separation in this thesis is thus essential.

4. Evaluation across multiple benchmark datasets with varying recording conditions (UCF101, PennAction, HAA500) under identical feature extraction and classification pipelines is largely absent from the sports HAR literature.

5. Statistical reliability analysis through multi-seed evaluation and bootstrap confidence intervals is rarely reported in lightweight HAR works, making it difficult to assess the stability and reproducibility of reported results.

CHAPTER 3

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR SPORTS POSE-BASED HUMAN ACTIVITY RECOGNITION

This chapter presents a comparative study of five classical machine learning models SVM, Random Forest, KNN, Logistic Regression, and XGBoost for sports-based Human Activity Recognition using MediaPipe extracted skeletal keypoints. A compact 396-dimensional statistical feature representation is derived from 33 body landmarks and evaluated across three benchmark datasets: UCF101, PennAction, and HAA500. The study demonstrates that SVM with RBF kernel consistently achieves the best performance, while confirming that lightweight pose-based descriptors combined with classical ML models offer an efficient and privacy-preserving alternative to computationally expensive deep learning approaches.

3.1 Introduction

HAR is very crucial in developing intelligent systems in the area of sports analytics, fitness monitoring, and rehabilitation assessment. Vision based motion analysis on whole videos is a very popular approach, and it works quite well under varying circumstances. However, but it has disadvantages of high computational cost and immense data requirements. Considering this, we study pose-based Human Activity Recognition, which deals with classifying skeletons key points into poses. This approach captures the structural movement of the human body using a set of joint coordinates, thus enabling robust and efficient activity recognition. It also significantly reduces computational load and data requirements. Additionally, this paradigm ensures privacy. Human motion can be analyzed using keypoints, without storing any identifiable visual data. This ability of recognition and classification of various athletic motions, such as pushups, pullups, squats, and jumping jacks, etc, is very critical in monitoring and analyzing the performance of athletes[10]. The ability to classify various motions enables automated coaching, performance analysis, form analysis, and injury prevention. As mentioned, traditional HAR systems based on CNNs trained on full-frame videos require high computational resources, thus impractical for edge deployment in wearable devices or gym-based sensors, where the analysis based on keypoint can work with ease.

The 2D human pose estimation frameworks, such as MediaPipe [33, 6], OpenPose[9], and AlphaPose [16], are very efficient for extracting 2D human skeletons in real time on

even low-end hardware. The Pose representations generated by these models result in a structured form that depicts human body dynamics in a very concise manner, due to which they are preferred when developing lightweight pose estimation systems. Building upon the efficiency of such works, there have been various efforts on both Machine Learning algorithms and lightweight Deep Learning architectures for recognition based on poses. Though ML models such as SVM, RF, and KNN are computationally efficient and perform well on small datasets, recent DL models are more capable of learning non-linear spatial temporal patterns in movement data [28, 42, 44].

There are still modern approaches other than MediaPipe BlazePose such as , RTMPose which can break through the established academic bias that top-down approaches must sacrifice speed. Top-down approaches can have both high accuracy and real-time speeds [23]. Following this trajectory, YOLOv8-Pose has emerged as a highly accurate convolutional neural network architecture that unifies object bounding box detection with keypoint estimation [54]. This unified architecture has drastically improved computational efficiency by reducing the number of model parameters by 13.63% and floating-point operations by 14.03% compared to previous baseline models [31].

Although several works focus on the HAR problem, only a few provide a side-by-side comparison between various machine learning algorithms. Most previous works either focused on large-scale general activities or used complex architectures that are unsuitable for deployment in low-power or wearable IoT devices. Few works investigated the trade-off between model complexity and performance across several datasets in the sports domain. To fill these gaps, this work presents a comparative analysis of machine learning approaches leveraging MediaPipe extracted pose keypoints. Evaluations have been carried out on three benchmark datasets: UCF101, PennAction, and HAA500.

3.2 Related Work

HAR has greatly advanced thanks to advances in computer vision and machine learning. Traditional methods of HAR used hand-crafted motion features computed from RGB or depth videos (dense trajectories, optical flows). Deep Learning models such as 2D/3D Convolutional Neural Networks (CNNs), RNNs, have proven successful in video HAR, however, they usually require large-scale annotated datasets, high computational cost, and time-consuming training procedures. As a result, deep learning models are not suitable for resource-constrained sports applications or edge devices. Also, using DL models trained on RGB video can be an issue regarding privacy as well as dealing with complex backgrounds and lighting conditions

[38]. Pose-based HAR has become increasingly popular due to its efficiency and privacy protection. The main idea of pose-based action recognition is to represent human motions not by the whole video frame but through a set of 2D or 3D keypoints, representing coordinates of certain body joints. Nowadays, frameworks such as MediaPipe[33, 6], OpenPose[9] and AlphaPose [16] provide real-time and accurate skeleton estimation on

commodity hardware. It makes it possible to apply ML algorithms to classify actions using interpretive pose features instead of raw images [59].

3.2.1 HAR in Sports

In sports, HAR is indispensable when assessing exercise quality, recognizing bad posture, and helping train athletes. Early studies mostly employed wearable sensors or RGB motion tracking, but nowadays, pose-based approaches are becoming more prevalent due to their efficiency and privacy concerns. It was proven that pose representation makes the algorithm more robust to changes in background and enables focusing only on body kinematics. Squats, push-ups, lunges, and jumping jacks are examples of sports actions that are well suited for pose representation since they have specific kinematic characteristics [10].

Apart from classification, modern systems tend to monitor exercises in real time and provide biomechanical feedback. Recently, it became possible to perform real-time repetition counting and exercise form recognition through the combination of MediaPipe and OpenCV [39, 50]. In addition, SmartFit makes use of this approach to help users exercise better [4]. All this work is based on the foundations laid by the algorithms like POGARS and Skeletor, where the superior robustness of pose-only representations was first proven [2, 22, 55].

Nevertheless, all the current research mainly deals with single sports datasets or exercise types, lacking evaluation across multiple sports datasets [8].

3.2.2 Machine Learning for Pose-Based HAR

There has been a considerable success in using ML classifiers for skeleton-based HAR owing to the simplicity, interpretability, and efficiency of such models [64]. SVMs are popular because of their capacity to deal with high-dimensional data and obtain accurate discrimination between different classes of activities. The RF model is known for its noise resistance and good performance in cases of heterogeneous features. KNN works well when solving small-scale problems where there are clearly defined clusters. The LR model provides a good benchmarking result in terms of linear classification. The ensemble methods like XGBoost are superior performers because of their ability to capture non-linear relations and reduce overfitting through gradient boosting [11].

3.3 Methodology

This section discusses the architecture used for the training and evaluation processes of our various ML models selected in our comparative research. The overall architecture

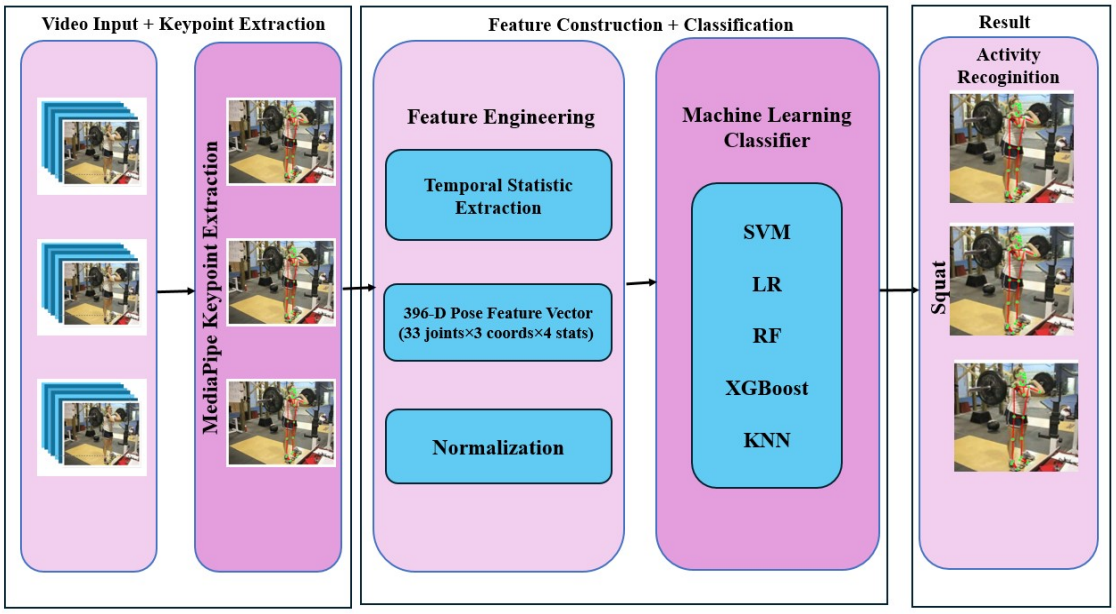

Figure 3.1: Overview of the proposed pose-based HAR pipeline including MediaPipe pose extraction, feature engineering, and ML-based classification

can be described in various phases which are: pose estimation using MediaPipe, feature engineering and normalization, and classification using Machine Learning. The overall architecture is shown in Fig. 3.1. It was uniformly applied to all three datasets UCF101 [45], PennAction [62], and HAA500 [14] and was limited to five classes of activity related to sports or exercise in each dataset.

3.3.1 Preprocessing

The preprocessing is divided into various stages as explained below. Here, each video sample undergoes the following stages:

Stage 1- Pose Estimation: Individual frames were processed using MediaPipe Pose [33, 6], which detects 33 body keypoints (x, y, z coordinates) as shown in Fig. 4.1. The model estimates joint positions such as the shoulders, elbows, hips, knees, ankles, and other anatomical points essential for describing athletic movements. The HAA500 dataset contains videos instead of frames so we directly processed the videos at their native frame rate with approx. 25–30 FPS, which also helped in maintaining temporal consistency, allowing the model to capture detailed motion patterns without downsampling. Frames where pose detection was not reliable i.e., missing keypoints were discarded for ensuring consistent skeleton information across clips. This affected less than 5% clips so overall it didn't had any major impact on the result.

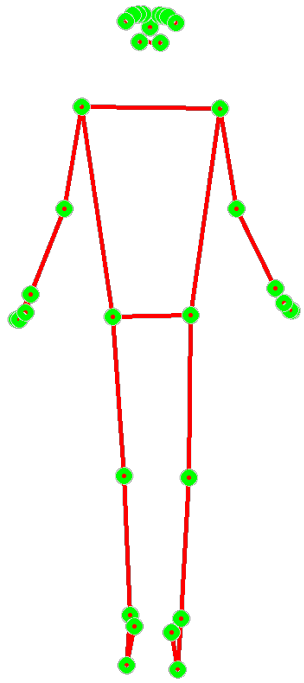

Figure 3.2: MediaPipe 33-Keypoint Skeleton Representation

Stage 2- Feature Construction: For each video sequence, the 2D pose coordinates (x, y), and relative depth (z) were aggregated across all frames. Since temporal sequence modeling was not utilized, we used some statistical descriptors such as mean, standard deviation, minimum, and maximum, which were computed for each joint coordinate, resulting in a feature representation of 396 dimensions (33 keypoints × 3 coordinates × 4 descriptors).
Stage 3- Feature Normalization: All features were normalized using StandardScaler to ensure uniform scaling and reduce inter-subject variation due to differences in camera position or participant size or distance of camera from subject.
Stage 4- Model Training and Evaluation: The normalized features and their corresponding activity labels were used to train all ML models. A 5-fold stratified cross-validation scheme per dataset was used for model evaluation, ensuring balanced class distribution in each fold. All models were trained using fixed hyperparameters to ensure a fair and consistent comparison.

3.3.2 Datasets and Selected Sports Classes

Three benchmark datasets available in public domain were considered for training and testing purposes, namely, UCF101 [45], PennAction [62], and HAA500 [14]. From each dataset, five sports-related action classes were chosen so as to keep uniformity in the nature

of actions involved. The classes were chosen in such a way that each dataset offers well-balanced representation of sports. The selected classes details are shown in Table 3.1.

Table 3.1: Selected sports activity classes from benchmark datasets

Dataset	Selected Classes	Samples
UCF101	Body Weight Squats (112), Jumping Jack (123), Lunges (127), PullUps (100), PushUps (102)	564
PennAction	Jumping Jacks (112), PullUp (199), PushUp (211), SitUp (100), Squat (231)	853
HAA500	Jumping_Jack (20), PullUps (20), PushUps (20), SitUp (20), Lunge (20)	100

3.3.3 Pose Estimation and Feature Representation

MediaPipe Pose [33, 6] was used to extract human skeletal keypoints due to its accuracy and real-time performance. For each frame t, MediaPipe provides 33 pose landmarks, and each landmark i is represented as a 3-dimensional vector:

$$\mathbf{p}_i^{(t)} = (x_i^{(t)},\, y_i^{(t)},\, z_i^{(t)}) \tag{3.1}$$

where (x, y) denote normalized 2D coordinates, z is a relative depth offset (2.5D representation). Since z is estimated from monocular RGB input, these values do not represent true 3D world coordinates. Given a video clip with T frames, the raw pose sequence is:

$$\mathbf{P} = \left\{ \mathbf{p}_i^{(t)} \;\middle|\; i = 1, \ldots, 33;\; t = 1, \ldots, T \right\}. \tag{3.2}$$

To obtain a compact, temporally invariant representation, we compute four statistical descriptors over time for each landmark attribute: mean, standard deviation, minimum, and maximum. For each landmark i and each attribute $k \in \{x, y, z\}$, these statistics are defined as follows:

$$\text{Mean:} \quad \mu_{i,k} = \frac{1}{T} \sum_{t=1}^{T} p_{i,k}^{(t)} \tag{3.3}$$

$$\text{Standard deviation:} \quad \sigma_{i,k} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(p_{i,k}^{(t)} - \mu_{i,k} \right)^2} \tag{3.4}$$

These descriptors are concatenated to construct the final feature vector, and each of the 33 landmarks contributes 3 attributes and 4 statistical operations. Thus, the resulting

feature dimensionality is:

$$\mathbf{f} = \left[\mu_{i,k},\ \sigma_{i,k},\ \min_{i,k},\ \max_{i,k}\right]_{i=1..33,\ k\in\{x,y,z\}}. \tag{3.5}$$

$$\dim(\mathbf{f}) = 33 \times 3 \times 4 = 396 \tag{3.6}$$

Thus, each video clip is represented as a fixed-length, 396-dimensional pose-based descriptor that captures key spatial and motion characteristics without requiring recurrent or sequence-based temporal models.

3.3.4 Machine Learning Models

We used five ML models in this work to classify sports activities according to pose features.

1. Support Vector Machines (SVM): Employed with the Radial Basis Function kernel in order to achieve the ability to capture non-linear boundaries between classes of activity allowing a balance of separation in the pose feature space.

2. Random Forest (RF): An ensemble consisting of two hundred decision trees, useful for understanding the importance of features, and helping to limit the consequence of overfitting through randomization from multiple weak learners.

3. K-Nearest Neighbors (KNN): A non-parametric baseline classifier that defines a class label based on majority vote within the k-nearest neighbors.

4. Logistic Regression (LR): Introduced with L2 regularization as a lightweight linear baseline to understand the separability of the pose-based feature distributions.

5. XGBoost: An ensemble boosting algorithm that builds on top of previously weaker learners is efficient in its design to minimize classification error sequentially [11]. Within the pose features, it allowed a capture of non-linear relationships while limiting overfitting.

3.3.5 Evaluation Metrics

The model performance was quantified using four metrics: Accuracy (Acc), Precision (P), Recall (R), and F1-Score (F1). Each model was evaluated using 5-fold stratified cross-validation, and the metrics (Accuracy, Precision, Recall, and F1-score) were averaged across the five folds.

3.3.6 Implementation Details

All experiments were implemented in Python 3.12 using Scikit-learn for ML models. MediaPipe Pose was used for keypoint detection, and NumPy along with Pandas handled data preprocessing. The training was done on Google Colab with a standard GPU runtime.

3.4 Results Analysis

This section presents a comparative evaluation of five ML models on three sports-oriented datasets: UCF101, PennAction, and HAA500. Each experiment used the same 396-dimensional pose-based feature representation described in Section 3.3 .

3.4.1 Results on UCF101

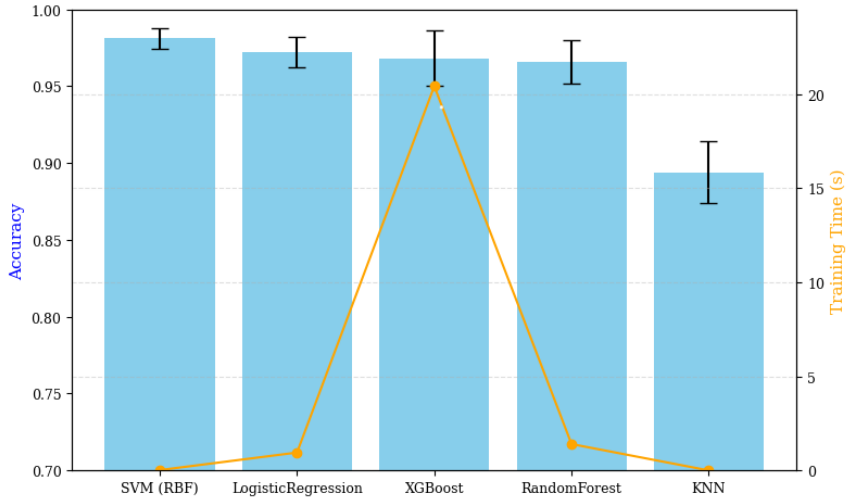

Figure 3.3: Performance comparison on UCF101 dataset

SVM (RBF) achieved the best performance on UCF101 with 98.1% accuracy and the highest precision/recall, showing strong generalization on the pose features. LR and XGBoost also performed competitively above 96%, while KNN remained the weakest due to its sensitivity to feature-scale variations. The detailed results are presented in Table 3.2 and Fig. 3.3 illustrates the accuracy and training-time comparison of the models.

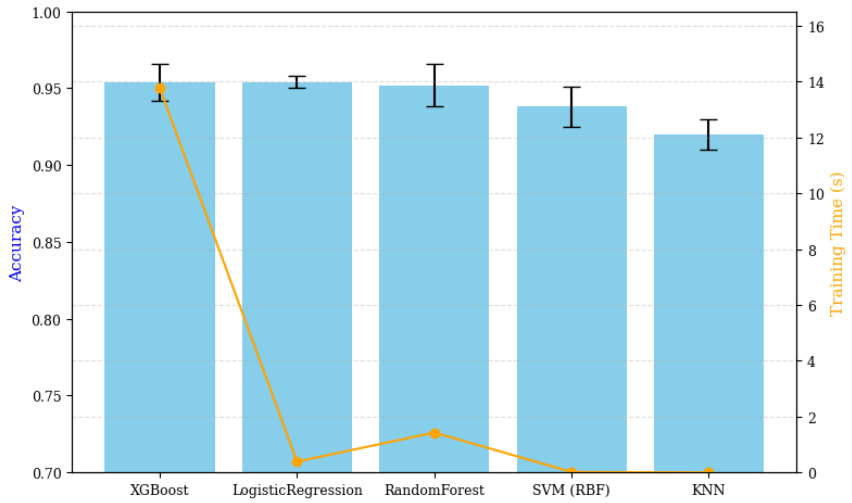

Figure 3.4: Performance comparison on PennAction dataset

3.4.2 Results on PennAction

As shown in Fig. 3.4, the performance of the ML models varies across the PennAction dataset. Both XGBoost and Logistic Regression achieved the highest accuracy of 95.4%, with XGBoost showing slightly higher recall. Random Forest also performed well (95.2%), while SVM and KNN lagged slightly behind. Table 3.2 summarizes the complete evaluation.

3.4.3 Results on HAA500

HAA500 proved to be the most challenging dataset, yet SVM (RBF) still achieved the highest accuracy (93%) and F1-score. Logistic Regression and XGBoost also performed strongly, while Random Forest and KNN showed comparatively lower performance. The full results are shown in Table 3.2 and Fig. 3.5 shows the model comparison on HAA500.

3.4.4 Overall Performance Analysis

As shown in Fig. 3.6 the comparative study reveals that SVM (RBF) provides the most reliable performance across heterogeneous sports datasets, achieving top accuracy on UCF101 (98.1%) and HAA500 (93%). Logistic Regression and XGBoost also showed

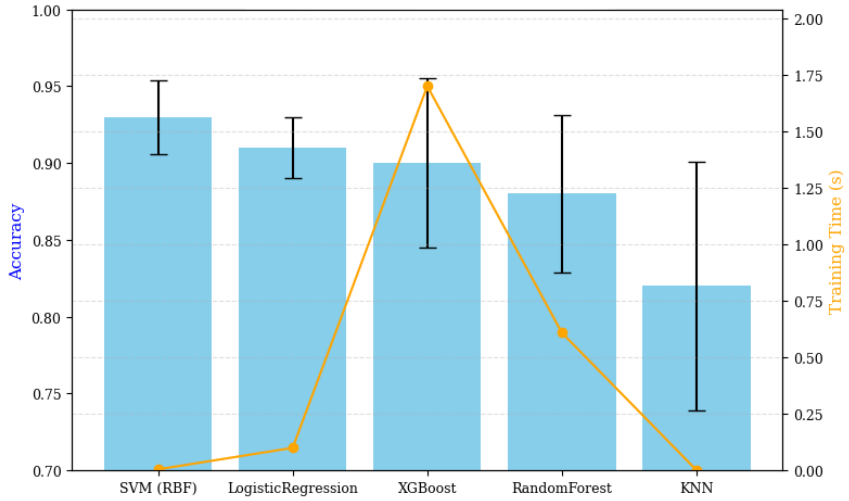

Figure 3.5: Performance comparison on HAA500 dataset

Table 3.2: Performance of ML models across all datasets (5-fold CV)

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Train Time (s)
UCF101	KNN	0.894	0.900	0.894	0.895	0.001
	Logistic Regression	0.972	0.974	0.971	0.972	0.953
	Random Forest	0.966	0.967	0.967	0.967	1.403
	SVM (RBF)	**0.981**	**0.981**	**0.981**	**0.980**	0.020
	XGBoost	0.968	0.972	0.968	0.969	20.425
PennAction	KNN	0.920	0.925	0.914	0.916	0.001
	Logistic Regression	**0.954**	**0.955**	0.949	0.950	0.387
	Random Forest	0.952	0.950	0.949	0.947	1.430
	SVM (RBF)	0.938	0.939	0.937	0.934	0.018
	XGBoost	**0.954**	0.953	**0.951**	**0.950**	13.761
HAA500	KNN	0.82	0.860	0.82	0.812	0.001
	Logistic Regression	0.91	0.929	0.91	0.909	0.100
	XGBoost	0.90	0.914	0.90	0.900	1.700
	Random Forest	0.88	0.902	0.88	0.881	0.609
	SVM (RBF)	**0.93**	**0.949**	**0.93**	**0.931**	0.005

competitive results, particularly on PennAction, where both reached 95.4% accuracy. Ensemble methods demonstrated strong recall due to their ability to capture non-linear variations in pose features. KNN performed the weakest, reflecting its sensitivity to high-dimensional spaces. Since inference only relies on precomputed 396-dimensional pose features and MediaPipe runs in real time on CPU/GPU, the entire pipeline can operate close to real–time on standard edge hardware.

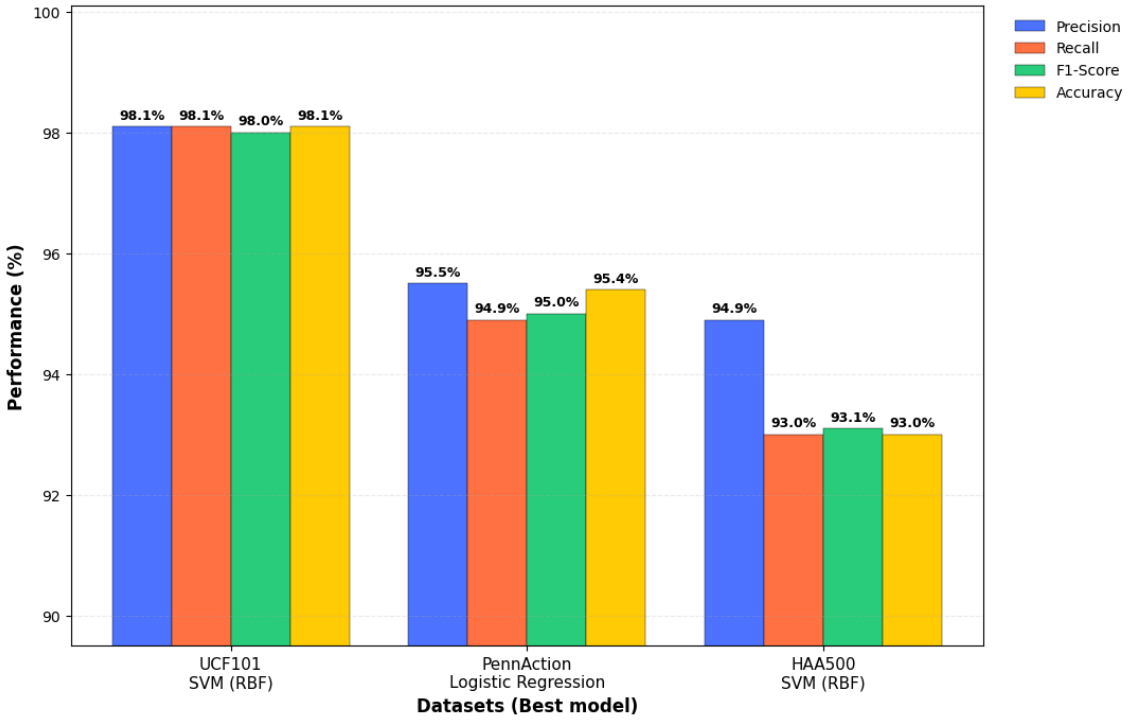

Figure 3.6: Performance comparison of the best ML models across the datasets

3.5 Summary

On the three sports-oriented benchmarks: UCF101, PennAction, and HAA500 , comparing the performances it clearly shows that SVM (RBF) offers the best performance with reliable overall performance for pose-based action recognition, reaching the highest accuracy on both UCF101 (98.1%) and HAA500 (93%), while Logistic Regression and XGBoost performed comparably, especially on PennAction. where both reached 95.4%. All models had more than 80% average accuracy, thereby confirming that skeletal descriptors alone contain sufficiently rich motion cues for distinguishing sports actions. Despite the strong results, the approach has certain limitations: temporal dynamics are reduced to simple statistical summaries, pose only inputs lack contextual information from RGB frames, and MediaPipe provides only 2.5D joint estimates. These factors may restrict performance in highly complex or ambiguous actions, especially when multiple sub-actions overlap temporally or when transitions between exercises occur without clear pose boundaries.

Further research will involve the use of LSTMs for enhancing temporal modelling as seen in the recent spatiotemporal LSTM-based HAR systems [42, 44]. Whether GRUs or transformer models are used, they will be able to detect more precise motion dynamics within consecutive frames. Other future areas of improvement could include 3D pose estimation and multi-modal data fusion in more difficult applications such as multi-person

pose estimation in team sports. The use of wearable IMUs and inertial units is another way to enhance performance in occluded scenes and complex body interactions in athletics.

The proposed pose-based approach supports privacy preserving activity monitoring by avoiding the use of identifiable RGB video data. Its lightweight computation makes realtime feedback feasible on low cost devices, enabling accessible fitness guidance and home-based rehabilitation. These aspects demonstrate practical value and potential social benefit in applications in sports, healthcare, and wellness.

CHAPTER 4

LIGHTWEIGHT POSE-BASED MULTI-CLASS EXERCISE RECOGNITION USING PCA-COMPRESSED SKELETAL FEATURES

This Chapter presents a lightweight pose-based exercise recognition method using PCA-compressed skeletal features instead of RGB data. It combines engineered motion features with a simple neural network to achieve high accuracy while reducing computational complexity. The approach demonstrates efficient and reliable performance on benchmark datasets, making it suitable for real-time applications.

4.1 Introduction

Human Activity Recognition (HAR) has become a fundamental problem in the computer vision field because it can be employed in tasks such as smart surveillance, fitness monitoring, rehabilitation, and human-computer interfaces [19, 20, 42]. The latest advancements in deep learning have also demonstrated very high performance based on RGB convolutional and transformer networks such as TimeSformer and Video Swin Transformer [7, 32]. Despite this success, these models are very computationally intensive, require large-scale data sets as well as prone to background clutter, change of illumination, and point of view. Skeleton-based HAR addressing these limitations is based on the concept of modeling motion with human pose keypoints rather than uncoded pixel values [30]. Graph Convolutional Networks (GCNs) form the basis for spatial-temporal dependence on joint graphs, including CTR-GCN and PoseC3D [12, 43, 15, 56]. However, they usually need complicated graph representation and a large amount of parameters [63], which may be unnecessary with organized and biomechanically different processes such as exercise movements. Recent HAR research has explored efficient transformer variants and lightweight graph-based models to reduce computational burden. However, many such methods still require complex architectures, large-scale training data, or specialized optimization. In contrast, the proposed method relies on interpretable engineered skeletal dynamics and dimensionality reduction, making it suitable for resource-constrained deployment scenarios.

Pose-based pipelines have gotten smaller and are now of interest with an increasing interest in deployable and lightweight systems [37]. It is possible to effectively carry out on-device keypoint detection with BlazePose [6] frameworks that detect pose keypoint

in real time. The former article conjectures that engineered skeleton attributes as well as shallow classifier and dimensionality shrinkage, can produce competitive performance and reduce the complexity of the model [62, 37]. In addition, sliding-window HAR may be susceptible to inappropriate window-level segmentation that introduces data leakage and inflated accuracy, which explains why there is a need to have rigorous video-level assessment procedures [63].

The work will describe an uncomplicated pose-based multi-class exercise recognition system, using skeletal movements compressed with PCA. It takes into account a five-class subset of the UCF101 dataset, who are JumpingJack, PullUp, PushUp, BodyWeightSquats and Lunges [45]. Further, a five-category subset of PennAction dataset [62] is also considered and they are squat, pushup, pullUp, JumpingJack and BenchPress. Human poses are extracted using BlazePose [6] and a small 13-joint configuration is employed to make a significant tradeoff between discriminative power and a tradeoff between computational efficiency.

Frame-level features are hybrid representations constructed by combining spatial joint coordinates, temporal motion features, joint angles, and limb distances. Sliding windows (30 frames with a stride of 15) are tallied to form a 2040 dimensional presentation which is subsequently reduced to 120 dimensions by Principal Component Analysis (95% of the variance preserved). The neural network (FFNN) classifier is trained under a rigid protocol of video-level division in an attempt to prevent data leakage.

On the UCF101 subset, the proposed model achieves 96.0% window-level accuracy and 98.25% video-level accuracy using majority voting. On the PennAction dataset, the model achieves 95.5% window-level accuracy and 95.0% video-level accuracy. Such results suggest that powerful multi-class exercise recognition can be achieved through the well-designed and compressed skeletal models, with only fractions of the architectural complexity of graph-based heavy models.

The key novelty of this work lies in demonstrating that compact skeletal representations combined with PCA-based temporal compression and a shallow feed-forward classifier can provide competitive recognition performance without requiring heavy graph convolutional or transformer architectures.

4.2 Methodology (Pipeline and Skeleton Representation)

4.2.1 Skeleton Generation

MediaPipe BlazePose was used to detect human poses in every single frame of each video and offers 2D keypoint detection in real-time. The BlazePose [6] produces 33 body landmarks, but to cut down on redundancy, a 13-joint body part set was used, but still provides discriminative motion information.

Joints that have been chosen are: neck (calculated as the mid point of the shoulders),

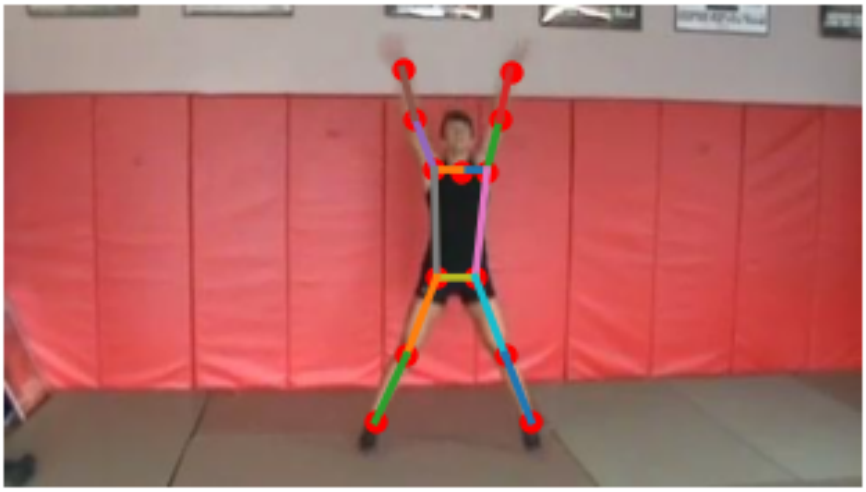

Figure 4.1: Selected 13-joint skeleton configuration

left and right shoulders, elbows, wrists, hips, knees and ankles. This arrangement is able to capture the necessary upper- and lower-body dynamics, and leave out facial and fine-scale hand landmarks, which are less informative to exercise recognition.

The selected 13-joint subset covers the principal upper-body and lower-body articulations involved in the considered exercises, including shoulders, elbows, wrists, hips, knees, and ankles. These joints capture the dominant motion patterns of push-ups, pull-ups, squats, lunges, and jumping jacks, while excluding facial and fine-grained landmarks that contribute limited discriminative value and increase feature dimensionality.

Each frame is represented as:

$$13 \text{ joints} \times 2 \text{ coordinates} = 26 \text{ spatial features} \tag{4.1}$$

4.2.2 Skeleton Preprocessing

The normalization was done in two phases to guarantee the invariance to camera translation and subject-scale.

Translation normalization: The coordinates of all the joints were transformed in a way that the neck joint was the origin. This eliminates unintentional spatial imbalances due to the position of the camera.

Scale normalization: The average length of torso (between neck and hips) was calculated as a scaling factor. They divided all the coordinates by this factor in order to normalize the size of the subject and the distance of the camera. This preprocessing guarantees spatial consistency in the videos shot with different viewpoints and proportions of the subject.

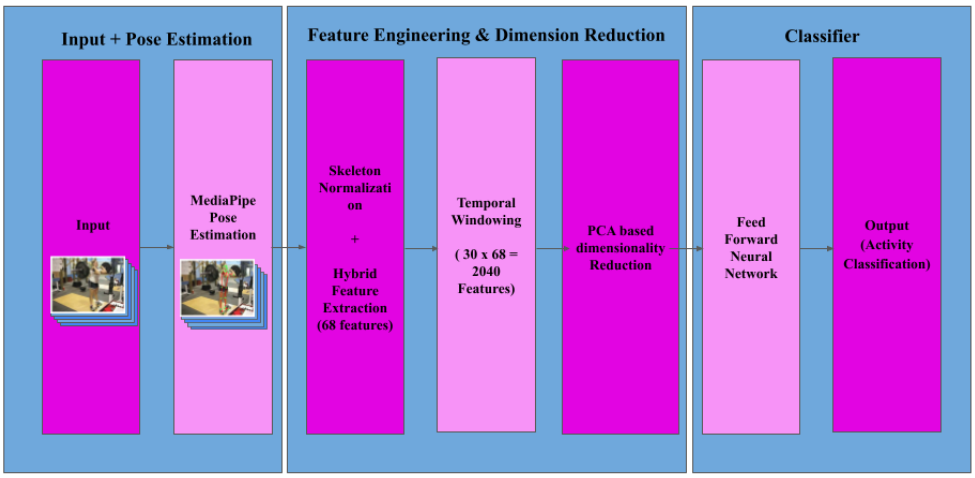

Figure 4.2: Overview of the proposed pose-based exercise recognition pipeline

4.2.3 Feature Extraction

A hybrid feature vector was generated using spatial, temporal, and biomechanical data of every normalized frame.

- **Spatial Coordinates:** 26 variables of normalized joint positions.
- **Motion Features:** Frrame-to-frame velocity computed from joint coordinate differences. This adds 26 motion features.
- **Joint Angles:** Angles of eight biomechanically relevant joints (elbows, knees, shoulders, hips) were calculated with vectors to determine an angle.
- **Limb Distances:** Structural relationships were represented by computing eight distance features between anatomically joined joints.

  The total frame-level feature dimension is:

$$26 \text{ (spatial)} + 26 \text{ (motion)} + 8 \text{ (angles)} + 8 \text{ (distances)} = 68 \text{ features} \tag{4.2}$$

4.2.4 Temporal Modeling

In order to model short-term dynamics of the motion, overlapping sliding windows were used with the window length equal to 30 frames and the stride length equal to 15 frames. Each window is a feature vector of:

$$30 \times 68 = 2040 \text{ features} \tag{4.3}$$

This approach captures temporal context as well as maintains lower computational cost.

4.2.5 Dimensionality Reduction

In order to eliminate the redundancy within the 2040-dimensional window representation, the Principal Component Analysis (PCA) was employed. PCA was only conducted on the training data with a 95% variance retention value. The reason for selecting 95% variance retention was to achieve a good compromise between retaining informative motion structures and eliminating redundancies. Through the experiments, the dimensionality reduction achieved through this configuration reduced the 2040-dimensional temporal representation into 120 and 91 dimensions for UCF101 and PennAction respectively, without significantly compromising recognition accuracy, suggesting that the technique achieved good feature compaction. The same fitted PCA was then used for transforming the testing data in order to avoid data leaking while significantly reducing computational power.

4.2.6 Classification Model

The final classifier was a lightweight FFNN. The network takes the feature vector of the PCA-compression as input and has three fully connected hidden layers with 256, 128 and 64 neurons respectively. The activation of every hidden layer is that of ReLU, dropout at a rate of 0.3 is used to enhance regularisation and minimise overfitting. The last output layer has five neurons that represent exercise classes. The remaining parameters that were used to train the model were the CrossEntropy loss function and the Adam optimiser with a learning rate of 0.001. The batch size was 64 and the maximum number of epochs was 30. To avoid overfitting, early stopping was used with the patience of five epochs, which was decided upon using validation performance.

4.3 Experimental Results

4.3.1 Datasets

Two benchmark human action datasets were experimented to determine performance on multiple datasets. The number of classes used was five out of the 101 in the UCF101, and it included Lunges, PushUp, BodyWeightSquats, JumpingJack, and PullUp. A strict video-level 80/20 split was used to prevent window-level data leakage, which resulted in 3040

training windows, 850 test windows and 114 distinct test videos. The lightweight, pose-based pipeline was loaded to a five-class subset of PennAction, composed of JumpingJack, PullUp, PushUp, Squat, and BenchPress. The same 13-joint pose model, normalization, hybrid feature extraction, windowing (30 frames, stride 15) and PCA compression (95% variance retention) were used. Video-level 80/20 stratified cut was used strictly.

4.3.2 Implementation Details and Evaluation Metrics

The framerwork was implemented using PyTorch. To prevent data leakage PCA was fitted exclusively on training data and applied to test data. The 2040-dimensional temporal features were reduced to 120 dimensions for UCF101 and 91 dimensions for PennAction while retaining 95% variance after using PCA. The model was trained using the Adam optimizer (learning rate = 0.001), with a batch size of 64 and a maximum of 30 epochs. Early stopping (patience = 5) based on validation performance was employed to prevent overfitting. Performance was evaluated at both window-level and video-level. At the window-level, standard classification metrics including Accuracy, Precision, Recall, F1-score, confusion matrix, ROC, and PR curves were used. Majority Voting [25] was used for final predictions for video level across all windows of that video.

4.4 Result Analysis

The proposed method works robustly on both data-sets, 96% acc (F1 = 0.96) on UCF101 and 95.5% (F1=0.95) on PennAction. For statistical evaluation, five random seeds are tested for proposed framework. The window level acc is 96.77% ± 0.22% and the video level acc is 98.09% ± 0.85%. Also the 95% confidence interval were [96.50%, 97.03%] and [97.03%, 99.14%] respectively, which shows robustness and stability over various runs.

Additionally, bootstrap resampling of the test predictions produced a window-level 95% confidence interval of $[93.86\%, 96.74\%]$, further supporting the robustness of the reported results.

4.4.1 Ablation Study

To validate the contribution of individual components, an ablation study was conducted on UCF101 dataset as presented in Table 4.1. The results show that PCA and temporal motion features are key contributors to performance. Removing PCA reduces accuracy, while excluding motion features causes the largest degradation, highlighting the importance of temporal dynamics. The proposed FFNN outperforms classical baselines and achieves competitive performance with LSTM at lower computational cost.

Table 4.1: Ablation and Baseline Comparison (UCF101)

Method	Accuracy (%)
Proposed (Full Model)	96.0
w/o PCA	92.8
w/o Motion Features	90.5
w/o Joint Angles	91.7
w/o Limb Distances	92.3
SVM	88.6
Random Forest	90.2
LSTM	93.5

Table 4.2: Window-Level Classification Performance on UCF101 and PennAction

	UCF101				PennAction			
Class	P	R	F1	Sup	P	R	F1	Sup
JumpingJack	0.98	1.00	0.99	102	1.00	1.00	1.00	10
PullUp	0.97	0.97	0.97	155	1.00	0.97	0.99	110
PushUp	0.91	0.99	0.95	79	0.97	0.90	0.93	62
Squats / BodyWeightSquats	0.97	0.90	0.93	170	0.92	0.97	0.94	226
Lunges / BenchPress	0.95	0.95	0.95	344	0.91	0.86	0.89	103
Accuracy	0.96 (850)				0.95 (511)			

4.4.2 Window-Level Performance

The model achieves high precision and recall across both datasets as shown in Table 4.2. Minor confusion is observed between squats and lunges in UCF101 due to similar lower-body motion patterns, while in PennAction, confusion primarily occurs between PushUp and BenchPress, as well as among lower-body exercises due to similar pose configurations.

4.4.3 Video-Level Performance

Temporal aggregation via majority voting improves performance, achieving 98.25% (F1 = 0.98) on UCF101 and 95.0% (F1 = 0.95) on PennAction. The gain over window-level results indicates improved stability and reduced prediction noise as shown in Table 4.3.

Table 4.3: Video-Level Classification Performance on UCF101 and PennAction

	UCF101				PennAction			
Class	P	R	F1	Sup	P	R	F1	Sup
JumpingJack	1.00	1.00	1.00	20	1.00	1.00	1.00	10
PullUp	0.98	0.98	0.98	30	1.00	0.97	0.99	39
PushUp	0.97	0.96	0.96	18	1.00	0.89	0.94	35
Squats / BodyWeightSquats	0.97	0.95	0.96	25	0.92	0.96	0.94	46
Lunges / BenchPress	0.98	0.98	0.98	21	0.87	0.96	0.91	27
Accuracy	0.9825				0.95			

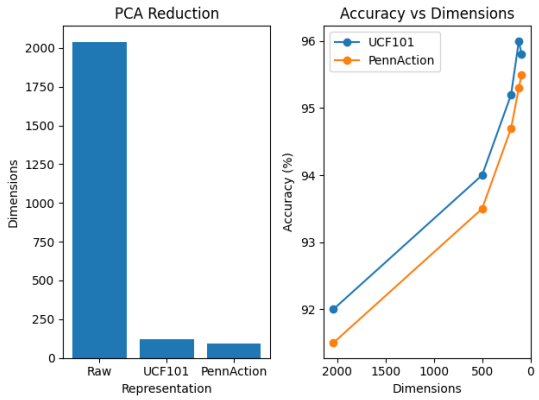

Figure 4.3: (Left) PCA dimensionality reduction from 2040 to 120/91. (Right) Accuracy vs. feature dimensions around 120 dimensions

4.4.4 Discussion

The robustness of the representation is verified by evaluations using two independent datasets. Even though UCF101 and PennAction differ in terms of recording conditions and performer variance, our framework is able to produce consistent results, demonstrating excellent generalization ability. As shown in Fig. 4.3, the use of PCA allows us to reduce the dimensionality of the features from 2040 down to 120 (UCF101) and 91 (PennAction), thus proving the presence of high redundancy in the temporal features. Moreover, the trend in accuracy clearly indicates that removing redundant features leads to improved results, achieving their maximum at around 120 dimensions and being stable afterwards, which highlights the usefulness of PCA.

Furthermore, we can see that our lightweight FFNN model outperforms more sophisticated spatio-temporal approaches, such as CTR-GCN and PoseC3D, while being computationally more efficient. Thus, the experiment shows that carefully crafted skeletal features along with feature dimensionality reduction can be considered as a feasible alternative to sophisticated spatio-temporal networks. Finally, the rigorous video-level splitting helps avoid data leakage in the experiments, resulting in reliable generalization ability.

4.5 Summary

This chapter presents a lightweight pose-based multi-class exercise recognition method using PCA-compressed skeletal dynamics. The method combines engineered spatial, temporal, and biomechanical features with dimensionality reduction and a shallow FFNN, avoiding computationally expensive graph-based architectures. It achieves 96% (window-level) and 98.25% (video-level) accuracy on UCF101, and 95.5% and 95.0% on PennAction. These results confirm that compact skeletal representations with PCA can achieve

strong recognition performance with low computational cost, making the approach suitable for real-time and resource-constrained environments. Future work will focus on scaling to larger datasets (e.g., NTU RGB+D), using 3D pose representations, and exploring lightweight temporal attention mechanisms for improved temporal modeling.

CHAPTER 5

RESULTS AND ANALYSIS

This chapter presents a comprehensive evaluation and comparative analysis of the two proposed approaches: (i) a machine learning-based pose classification framework using statistical skeletal features, and (ii) a lightweight deep learning approach using PCA-compressed skeletal dynamics. The first study is evaluated on 3 benchmark dataset UCF101, PennAction, HAA500 and the second study is evaluated on 2 benchmark datasets UCF101 and PennAction.

5.1 Overall Performance Comparison

Both methods illustrate the effectiveness of using pose representations in HAB with very little computational cost. The machine learning method proves to be very effective for many datasets, while the PCA method is almost equally accurate and more efficient.

Table 5.1: Overall Accuracy Comparison of Proposed Approaches

Dataset	ML-Based (Best Model)	PCA-Based (Window)	PCA-Based (Video)
UCF101	98.1% (SVM)	96.0%	98.25%
PennAction	95.4% (LR/XGBoost)	95.5%	95.0%

5.2 Dataset-wise Analysis

The ML-based method gives the best peak accuracy when it is at its best, whereas the PCA-based model has similar performances with less complex and a better temporal modeling. Both of the methods get very high accuracy on the UCF101 dataset. PennAction performs well in both approaches (with similar performance, around 95%), suggesting that pose-based features are discriminative enough, irrespective of different modeling strategies. With higher variability and complexity of actions, performance is slightly reduced on HAA500. But SVM remains very effective, thus showing the robustness of statistical pose features.

5.3 Model Performance Comparison

Two performance metrics *(accuracy, F1-score)* were used to assess the performance of different machine learning models on three benchmark datasets *(*UCF101, PennAction, HAA500). In general, Support Vector Machine (SVM) and Logistic Regression (LR) can be seen to have strong performance across the sets with highest performance of SVM in *UCF101 and HAA500*. LR and XG-Boost have similar and better performance on *PennAction* dataset. Random Forest (RF) and XGBoost are ensemble classifiers that perform well and have stable results, effectively capturing non-linear patterns. In comparison, KNN does not perform very well for any dataset, possibly because of its sensitivity to feature scaling and high-dimensional data. The results reveal the usefulness of linear and margin based classifiers in human pose-based activity recognition tasks.

Table 5.2: Performance of ML Models Across Datasets

Dataset	Model	Accuracy	F1-Score	Remarks
UCF101	SVM	98.1%	0.98	Best performer
	LR	97.2%	0.97	Strong baseline
	RF	96.6%	0.96	Stable performance
	XGBoost	96.8%	0.96	Good non-linear modeling
	KNN	89.4%	0.89	Sensitive to scaling
PennAction	LR	95.4%	0.95	Best performer
	XGBoost	95.4%	0.95	Comparable to LR
	RF	95.2%	0.94	Stable
	SVM	93.8%	0.93	Slight drop
	KNN	92.0%	0.91	Weakest
HAA500	SVM	93.0%	0.93	Best performer
	LR	91.0%	0.91	Strong baseline
	XGBoost	90.0%	0.90	Competitive
	RF	88.0%	0.88	Moderate
	KNN	82.0%	0.81	Weakest

5.4 Window-Level vs Video-Level Performance

A key advantage of the PCA-based approach is the temporal modeling using sliding windows and majority voting.

The results showcases that video-level aggregation improves stability and reduces prediction noise, hence improving overall accuracy of the model.

Table 5.3: Window-Level vs Video-Level Performance

Dataset	Window-Level Accuracy	Video-Level Accuracy
UCF101	96.0%	98.25%
PennAction	95.5%	95.0%

5.5 Feature Effectiveness and Ablation Analysis

The PCA approach highlights the importance of feature design and dimensionality reduction. PCA is used to reduce the number of features from 2040 to almost 120 without losing 95% of the variance, thus improving the computational efficiency.

Table 5.4: Ablation Study on UCF101 (PCA-Based Model)

Configuration	Accuracy
Full Model	96.0%
Without PCA	92.8%
Without Motion Features	90.5%
Without Joint Angles	91.7%
Without Limb Distances	92.3%

The results indicate that temporal motion features and PCA contribute positively into performance of model.

5.6 Computational Efficiency

Both approaches are computationally efficient compared to deep learning methods:

- ML-based approach uses fixed 396-dimensional feature vectors with fast training time.
- PCA-based approach reduces high-dimensional temporal features and uses a lightweight FFNN.

These characteristics make both methods suitable for real-time deployment on edge devices.

5.7 Key Observations

- The accuracy of Pose-based HAR is high and does not require RGB data.

- In terms of consistency across datasets, SVM is the most consistent out of all the ML models.
- The temporal modeling based on PCA enhances robustness and stability.
- In lightweight systems, feature engineering is a crucial component.
- Both methods are applicable to several datasets and have a good generalization property.

5.8 Conclusion of Analysis

The co-evaluation shows the efficiency and accuracy of both methodologies in the context of pose-based activity recognition. The PCA-based approach is more robust and improves the temporal modeling capabilities, whereas the ML models are easy to implement and have a good baseline performance. These results validate the viability of using the lightweight skeletal models to substitute for expensive deep learning models in real-world HAR scenarios.

5.9 Statistical Reliability

The robustness of the PCA-based model was assessed by running the model several times using different random seeds. The model showed a high mean accuracy of 96.77% with SD (standard deviation) of ±0.22% at window level, which means that the model has stable performance. Furthermore, the consistency of results was further confirmed by using confidence intervals and bootstrapping over various data splits, showing that the proposed approach generalizes well across splits.

5.10 Comparative Discussion

The comparison makes it clear that there is a trade off between simplicity and temporal modelling. The ML-based approach is easier to implement, is fast, and performs well without a high degree of complexity. But there is no explicit temporal modelling. The PCA-based approach, however, integrates temporal dynamics in the form of sliding windows, and dimensionality reduction, which yields greater robustness and stability at the video level.

In conclusion, both methods are successful, but the method based on PCA is more suitable for real-world applications due to its superior performance and scalability.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIETAL IMPACT

6.1 Conclusion

In this thesis, a detailed investigation of pose-based Human Activity Recognition (HAR) approaches was conducted focusing on lightweight algorithms that could be implemented in practice. The two approaches examined in this study include machine learning algorithms using statistical skeletal features, as well as PCA with compressed skeletal dynamics, which are modeled using a feed-forward neural network. The results of experiments conducted on various benchmarks, including UCF101, PennAction, and HAA500, showed that pose-based representations could be highly accurate even without the implementation of computationally expensive RGB models. Particularly, machine learning algorithms produced highly accurate results, and the most successful among those tested were Support Vector Machines. Meanwhile, the PCA-based approach proved to be highly robust thanks to temporal modeling and dimensionality reduction.

Thus, it is possible to state that pose-based approaches combined with light models could efficiently recognize human activities. Besides, PCA allows for considerable dimensionality reduction in feature space without any loss of discriminative power, which increases the efficiency of computations.

6.2 Future Scope

Despite the promising performance, there are certain aspects that could be improved in the system:

- **Temporal Deep Learning Models:** Application of sequence-based models like LSTM, GRU, and transformers would benefit from modeling temporal dependencies.
- **3D Pose Estimation:** The extension of the framework for the true 3D skeleton data will improve accuracy and give more information regarding the position of body parts. This would be particularly useful for difficult actions.
- **Multi-Person Activity Recognition:** Development of the approach towards recognition of several people at once will be useful for practical application such as in sports analysis and surveillance.

- **Multi-Modal Fusion:** Using other types of modalities along with pose estimation data like RGB images, depth, and IMUs will increase the robustness of the approach in challenging environmnets.
- **Real-Time Deployment:** Optimization of the system for deployment in real-time settings, including edge computing on devices like smartphones and embedded systems, and wearable devices such as fitness trackers.
- **Scaling to Large Data:** Evaluation of the proposed approaches on a large scale, for instance on the NTU RGB+D dataset.

6.3 Societal Impact

Proposed pose-based HAR systems have great potential in terms of positive societal impact:

- **Healthcare and Rehabilitation:** Monitoring of exercise using HAR can be helpful in rehabilitating patients, allowing them to perform their exercises autonomously without continuous supervision.
- **Fitness and Wellness:** Applications that provide intelligent guidance for performing exercises will help people avoid injuries and lead to improved health outcomes.
- **Sport Performance Analysis:** Athletes/trainers could make use of the system for analysis of movements in their bodies and improve their performance and decrease the probability of injury. Preserving
- **Privacy:** As opposed to prior systems that obtain visual features from RGB frames, the developed method does not take any images that can identify people, so it can be applied to privacy-aware application.
- **Accessibility:** Its light-weight design will allow powerful monitoring applications to be carried out even in low resource environments.

However, there are several aspects to be taken care of:

- **Generalization:** small amount of data can lead to over-fitting, and then performance on unseen data will deteriorate.
- **Ethical Considerations in applications:** the implementation of activities like monitoring using HAR technologies must be handled carefully.

In summary, the HAR can be utilized for many health and fitness related applications and will have a potential to help in building intelligent systems

BIBLIOGRAPHY

[1] Dustin Aganian et al. Including semantic information via word embeddings for skeleton-based action recognition. *arXiv preprint arXiv:2506.18721*, 2025.

[2] Anonymous. Pose is all you need: The pose only group activity recognition system (pogars). *arXiv preprint arXiv:2108.04186*, 2021.

[3] Anonymous. Lightweight graph convolutional network for efficient skeleton-based action recognition. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024.

[4] Anonymous. Smartfit: Real-time pose estimation and movement analysis. In *2025 International Conference on Computational Innovations and Engineering Sustainability (ICCIES)*. IEEE, 2025.

[5] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of the International Conference on Pervasive Computing*, pages 1–17, 2004.

[6] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. BlazePose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204, 2020.

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 813–824, 2021.

[8] Wenzheng Cai, Yuhang Luo, and Xin Yang. Pose-based human activity recognition using machine learning for fitness applications. *Heliyon*, 8(8):e10215, 2022.

[9] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. volume 43, pages 172–186, 2021.

[10] Chao Chen, John Daponte, and Guido Matrella. Estimating physical activities and exercises from skeletons using machine learning. *Sensors*, 20(2):588, 2020.

[11] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016.

[12] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13359–13368, 2021.

[13] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20186–20196, 2022.

[14] Jehoon Chung, Cheng-Han Wuu, Hsiao-Ru Yang, Yu-Wing Tai, and Chi-Keung Tang. HAA500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13465–13474, 2021.

[15] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition with pose-based 3D convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[16] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Jiefeng Li, and Cewu Lu. AlphaPose: Whole-body pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[17] Pedro Felzenszwalb and Daniel Huttenlocher. Pictorial structures for object recognition. volume 61, pages 55–79, 2005.

[18] Martin A. Fischler and Robert A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[19] Y. Guan and T. Ploetz. Ensembles of deep learning architectures for activity recognition: A review. *ACM Computing Surveys*, 2024.

[20] Y. Guo, J. Li, and Y. Wang. Skeleton-based human action recognition: A survey of deep learning approaches. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[21] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[22] S Jiang, B Sun, L Wang, Y Bai, K Li, and Y Fu. Skeletor: skeletal transformers for robust body-pose estimation. *PeerJ Computer Science*, 2021.

[23] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

[24] Jesus Jimenez, Joni McCawley, Francis A. Mendez Mediavilla, and Abhimanyu Sharotry. Advancements in human action recognition in manual material handling using machine learning. In *IMHRC 2025*, 2025.

[25] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[26] Khadija Lasri, Khalid El Fazazy, Adnane Mohamed Mahraz, Hamid Tairi, and Jamal Riffi. Lst-agcn: A novel unified lightweight attention framework for efficient skeleton-based action recognition. *Big Data and Cognitive Computing*, 10(4):125, 2026.

[27] Boyu Li et al. Real-time human activity recognition on edge microcontrollers: Dynamic hierarchical inference with multi-spectral sensor fusion. *Measurement Science and Technology*, Apr 2026.

[28] Yong Li, Chang Xu, and Jian Wang. Pose-based human action recognition using spatial–temporal joint encoding. *Multimedia Tools and Applications*, 82:33001–33022, 2023.

[29] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 29248–29257, 2025.

[30] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):3007–3021, 2020.

[31] Y Liu, Z Fu, L Song, H Sui, and B Li. Improved yolov8-pose algorithm for albacore tuna (thunnus alalunga) fork length extraction and weight estimation. *Journal of Marine Science and Engineering*, 12(5):784, 2024.

[32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, and Stephen Lin. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

[33] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, et al. MediaPipe: A framework for building perception pipelines. In *CVPR Workshops*, pages 1661–1664, 2019.

[34] Saakshi M V, Dhanya Rao, Nidhi M, Anurag Hurkadli, and Sneha Varur. Sportposenet: Leveraging pose estimation and deep learning for sports activity classification. In *Proceedings of the 3rd International Conference on Futuristic Technology - Volume 3: INCOFT*, pages 581–588. SciTePress, 2025.

[35] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

[36] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. volume 25, pages 24–38, 2014.

[37] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 223:103553, 2022.

[38] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. *Computers & Graphics*, 85:85–98, 2020.

[39] Riccardo Riccio. Real-time fitness exercise classification and counting from video frames. *arXiv preprint arXiv:2411.11548*, 2024.

[40] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 32–36, 2004.

[41] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[42] Abhilasha Sharma and Roshni Singh. ConvST-LSTM-Net: Convolutional spatiotemporal LSTM networks for skeleton-based human action recognition. *International Journal of Multimedia Information Retrieval*, 12(2):34, 2023.

[43] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Roshni Singh and Abhilasha Sharma. STAD-ConvBi-LSTM: Spatio-temporal attention-based deep convolutional bi-LSTM framework for abnormal activity recognition. *Journal of Visual Communication and Image Representation*, page 104465, 2025.

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5815–5824, 2019.

[47] Muhammad Tayyab, Sulaiman Abdullah Alateyah, Mohammed Alnusayri, Mohammed Alatiyyah, Dina Abdulaziz AlHammadi, Ahmad Jalal, and Hui Liu. A hybrid approach for sports activity recognition using key body descriptors and hybrid deep learning classifier. *Sensors*, 25(2):441, 2025.

[48] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014.

[49] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3D skeletons as points in a Lie group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 588–595, 2014.

[50] J Vijaya, Ashutosh Singh, Khushdeep Singh, and Amit Kumar. Ai-enabled real-time exercise monitoring with mediapipe and opencv. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024.

[51] Jingyao Wang, Emmanuel Bergeret, and Issam Falih. Skeleton-based action recognition with spatial-structural graph convolution. *arXiv preprint arXiv:2407.21525*, 2024.

[52] Bin Wu, Mei Xue, Ying Jia, Ning Zhang, GuoJin Zhao, XiuPing Wang, and Chunlei Zhang. Lightweight and efficient skeleton-based sports activity recognition with astm-net. *PLOS ONE*, 20(7):e0324605, 2025.

[53] Qingsheng Xie and Hongmin Deng. Li-agcn: A lightweight initialization-enhanced adaptive graph convolutional network for effective skeleton-based action recognition. *Sensors*, 25(23):7282, 2025.

[54] X Xu, T Wu, Z Du, H Rong, S Wang, S Li, et al. Yolov8-pose based behavior tracking of calves in precision livestock farming. *PLOS ONE*, May 2025.

[55] Zixuan Xu, Yichun Lou, Yang Song, and Zihuai Lin. Fitness action recognition based on mediapipe. *Procedia APA*, 2022.

[56] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[57] Tao Yu, Sandro Pinto, Tiago Gomes, Adriano Tavares, and Hao Xu. Clustered federated spatio-temporal graph attention networks for skeleton-based action recognition. *Sensors*, 25(23):7277, 2025.

[58] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2752–2759, 2013.

[59] Jian Zhang, Wei Liu, Pichao Wang, and Changxin Gao. A comprehensive survey on skeleton-based human activity recognition with deep learning. *Multimedia Tools and Applications*, 80(9):13411–13444, 2021.

[60] Qiujie Zhang, Wufan Wang, Bo Zhang, Zheng Zhang, Xirong Que, and Wendong Wang. A lightweight real-time framework for skeleton-based action recognition on mobile devices. In *IEEE Conference*, 2025.

[61] Weiyu Zhang, Menglong Zhu, and Konstantinos Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[62] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2624–2631, 2013.

[63] Xikun Zhang, Chang Xu, and Dacheng Tao. A comprehensive survey on skeleton-based action recognition with deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5410–5431, 2022.

[64] Lei Zhao and Han Zhang. Comparative study of machine learning algorithms for human activity recognition using pose-based features. *Pattern Analysis and Applications*, 25:1421–1435, 2022.

[65] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

LIST OF PUBLICATIONS AND THEIR PROOFS

Published/Communicated

INTERNATIONAL CONFERENCES

1. Aman Yadav, et al. "Comparative Analysis of Machine Learning Models for Sports Pose-Based Human Activity Recognition," *10th International Conference on Innovative Computing and Communication (ICICC-2027)*, Shaheed Sukhdev College of Business Studies, University of Delhi, India, in association with University of Valladolid, Spain. **(Accepted and presented; Springer LNNS Series, Scopus Indexed)**.

2. Aman Yadav, et al. "Lightweight Pose-Based Multi-Class Exercise Recognition Using PCA-Compressed Skeletal Features," *2nd Congress on Intelligent Machines and Algorithms (CIMA 2026)*, National Institute of Technology Puducherry, India. **(Accepted, To be Presented; Springer LNNS Series, Scopus Indexed)**.

DECLARATION

We hereby certify that the work which is presented in the Major Project-II/Research Work entitled **"Efficient Pose-based Human Activity Recognition Through Skeletal Feature Engineering and Dimensonality Reduction"** in fulfillment of the requirement for the award of the Degree of Master of Technology in **Software Engineering** and submitted to the Department of **Software Engineering**, Delhi Technological University, Delhi, is an authentic record of our own work, carried out during a period from **August 2024 to May 2026** under the supervision of **Dr. Abhilasha Sharma**.

The matter presented in this report/thesis has not been submitted by us for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/SCI-expanded/SSCI/Scopus indexed journal OR peer-reviewed Scopus indexed conference with the following details:

Title of the Paper: Comparative Analysis of Machine Learning models for pose-based Human Activity Recognition
Author Names (in sequence as per research paper): Aman Yadav, Abhilasha Sharma
Name of Conference/Journal: 9th INTERNATIONAL CONFERENCE ON INNOVATIVE COMPUTING AND COMMUNICATION - 2026
Conference Dates with venue (if applicable): 06-07 Feb 2026, Shaheed Sukhdev College of Business Studies, India
Have you registered for the conference?: Yes
Status of paper (Accepted/Published/Communicated): Accepted and Presented (Scopus-Indexed)
Date of paper Communication: 15 November 2025
Date of paper Acceptance: 9 December 2025
Date of paper Publication: To be published

24/SWE/12
AMAN YADAV
Student(s) Roll No., Name and Signature

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I further certify that the publication and indexing information given by the student is correct.

Place: Delhi
Date: July 7, 2026

Dr. Abhilasha Sharma
Supervisor Name and Signature

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (conference website or ScienceDirect in case of journal publication).

1. Aman Yadav, et al. "Comparative Analysis of Machine Learning Models for Sports Pose-Based Human Activity Recognition," 10th International Conference on Innovative Computing and Communication (ICICC-2027), Shaheed Sukhdev College of Business Studies, University of Delhi, India, in association with University of Valladolid, Spain. **Accepted and presented; Springer LNNS Series, Scopus Index.**

(ICICC - 2026)

INTERNATIONAL CONFERENCE ON INNOVATIVE COMPUTING AND COMMUNICATION (ICICC-2026)

Certificate

This is to certify that **Prof. / Dr. / Mr. / Ms.** **Aman Yadav** is a presenter/co-author of the paper titled **Comparative Analysis of Machine Learning Models for Sports Pose-Based Human Activity Recognition** in the **9th International Conference on Innovative Computing and Communication (ICICC-2026)** organized by Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, India in association with the National Institute of Technology Patna, India and the University of Valladolid, Spain on **06th-07th February 2026**.

Prof. Poonam Verma
Principal, SSCBS
University of Delhi,
New Delhi, India

Prof. Prabhat Kumar
General Chair
National Institute of
Technology, Patna, India

Prof. A.K. Singh
Technical Program Chair
National Institute of Technology,
Kurukshetra, Haryana, India

Presentation Certificate for ICICC-2026

ICICC 2026

The ninth edition of the International Conference on Innovative Computing and Communication (ICICC-2026) was successfully organized at Shaheed Sukhdev College of Business Studies, New Delhi, India, in association with the National Institute of Technology Patna, India, and University of Valladolid, Spain, on 06–07 February 2026.

The conference witnessed an overwhelming global response, receiving more than 2,500 research paper submissions from over 9,000 authors worldwide, reflecting the growing international interest in emerging computing and communication technologies. After a rigorous peer-review process, 475 high-quality papers were accepted, resulting in a competitive acceptance rate of approximately 18%.

All accepted papers were published in Springer's Lecture Notes in Networks and Systems, a Scopus-indexed publication series. The conference attracted over 1,000 participants including researchers, academicians, keynote speakers, industry experts, and delegates from around 40 countries, making ICICC-2026 a truly global platform for knowledge exchange and collaboration.

The event was conducted in hybrid format enabling both physical and virtual participation. The first day featured four distinguished keynote addresses delivered by eminent experts from prestigious institutions such as Indian Institute of Technology, Robert Morris University (USA), Saint Petersburg State University (Russia), and University of Southampton (Delhi Campus).

A major highlight of the conference was the Universal Innovators Leadership Awards (UILA) ceremony, recognizing outstanding contributions from industry leaders and innovators. The awards received around 750 nominations, with approximately 20% shortlisted and only 1–2% finally selected across categories.

Over the two-day conference, the technical program included five offline technical sessions and forty online sessions, providing a vibrant platform for presenting cutting-edge research, fostering academic discussions,

Important Dates

Submission of revised manuscript
20th November 2026

Notification of Acceptance / Rejection
30th November 2026

Final manuscript due
20th December 2026

Registration Deadline

Scopus Proof for ICICC-2026

DECLARATION

We hereby certify that the work which is presented in the Major Project-II/Research Work entitled **"Efficient Pose-based Human Activity Recognition Through Skeletal Feature Engineering and Dimensonality Reduction"** in fulfillment of the requirement for the award of the Degree of Master of Technology in **Software Engineering** and submitted to the Department of **Software Engineering**, Delhi Technological University, Delhi, is an authentic record of our own work, carried out during a period from **August 2024 to May 2026** under the supervision of **Dr. Abhilasha Sharma**.

The matter presented in this report/thesis has not been submitted by us for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/SCI-expanded/SSCI/Scopus indexed journal OR peer-reviewed Scopus indexed conference with the following details:

Title of the Paper: LightWeight Pose-Based Multi-class Exercise Recognition Using PCA Compressed Skeletal Feature
Author Names (in sequence as per research paper): Aman Yadav, Abhilasha Sharma
Name of Conference/Journal: 2nd Congress on Intelligent Machines and Algorithms (CIMA 2026)
Conference Dates with venue (if applicable): 23-24 May 2026, National Institute of Technology, Puducherry, India
Have you registered for the conference?: Yes
Status of paper (Accepted/Published/Communicated): Accepted and Presented (Scopus-Indexed)
Date of paper Communication: 20 April 2026
Date of paper Acceptance: 01 May 2026
Date of paper Publication: To be published

24/SWE/12
AMAN YADAV
Student(s) Roll No., Name and Signature

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I further certify that the publication and indexing information given by the student is correct.

Place: Delhi
Date: July 7, 2026

Dr. Abhilasha Sharma
Supervisor Name and Signature

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (conference website or ScienceDirect in case of journal publication).

2. Aman Yadav, et al. "Lightweight Pose-Based Multi-Class Exercise Recognition Using PCA-Compressed Skeletal Features," 2nd Congress on Intelligent Machines and Algorithms (CIMA 2026), National Institute of Technology Puducherry, India.

Presented; Springer LNNS Series, Scopus Indexed

2nd Congress on Intelligent Machines and Algorithms (CIMA 2026)

Organized by

National Institute of Technology, Puducherry, India

Certificate of Presentation

This is to certify that

Aman Yadav

has presented the paper title **Lightweight Pose-Based Multi-Class Exercise Recognition Using PCA-Compressed Skeletal Features** authored by **Aman Yadav, Abhilasha Sharma** in the **2nd Congress on Intelligent Machines and Algorithms (CIMA 2026)** held at **National Institute of Technology, Puducherry, India** during **May 23-24, 2026**.

SCRS/CIMA2026/PC/1393

Touhid Bhuiyan
Washington University of Science and Technology, USA
(General Chair)

Malaya Kumar Nath
National Institute of Technology Puducherry, Karaikal, India
(General Chair)

Springer

https://scrs.in/conference/CIMA2026

CIMA 2026 Presentation Certificate

CIMA 2026 Scopus Proof

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis _____

_____

Total Pages _____ **Name of the Scholar** _____

Supervisor(s)
(1) _____

(2) _____

(3) _____

Department _____

This is to report that the above thesis was scanned for similarity detection process and outcome is given below:

Software used: _____ Similarity Index: _____ Total Word Count: _____

Date: _____

Candidate's Signature **Signature of Supervisor(s)**

Aman Yadav

Aman_Pkka_final_version

PlagReportThesis

Document Details

Submission ID

trn:oid:::27535:140031585

Submission Date

May 22, 2026, 3:47 PM GMT+5:30

Download Date

May 25, 2026, 2:47 PM GMT+5:30

File Name

Aman_Pkka_final_version.pdf

File Size

4.8 MB

78 Pages

18,506 Words

111,295 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Aman Yadav

Aman_Pkka_final_version

PlagReportThesis

Document Details

Submission ID

trn:oid:::27535:140031585

Submission Date

May 22, 2026, 3:47 PM GMT+5:30

Download Date

May 25, 2026, 8:28 AM GMT+5:30

File Name

Aman_Pkka_final_version.pdf

File Size

4.8 MB

78 Pages

18,506 Words

111,295 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 10 words)

Match Groups

72 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

13 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
3% Publications
6% Submitted works (Student Papers)

AUTHOR'S BIOGRAPHY

Aman Yadav is currently pursuing M.Tech in Software Engineering from Delhi Technological University (DTU), New Delhi. He has obtained his Bachelor's degree in Computer Science and Engineering where he had acquired skills in programming, machine learning and computer vision. His research interests lie in the field of Human Activity Recognition (HAR), Computer Vision and Lightweight Deep Learning Models for real-world application. During his Master's degree program he had a good experience working on pose-based activity recognition using skeletal key-points with emphasis on developing computationally efficient model for edge deployment. He has also participated in research works within the domain of pose-based HAR including comparison of ML models and developing a PCA-based lightweight neural framework for exercise recognition.