

Mansi Gupta

Mansi_24/SWE/09



PlagReportThesis

Document Details

Submission ID

trn:oid:::27535:142861277

Submission Date

Jun 13, 2026, 6:41 PM GMT+5:30

Download Date

Jun 13, 2026, 6:48 PM GMT+5:30

File Name

Mansi_24_SWE_09.pdf

File Size

28.0 MB

118 Pages

27,979 Words

166,827 Characters

5% Overall Similarity

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

Filtered from the Report

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

Match Groups

-  **84 Not Cited or Quoted 5%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
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

- 2%  Internet sources
- 2%  Publications
- 4%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 84 Not Cited or Quoted 5%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
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

- 2% Internet sources
- 2% Publications
- 4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** Student papers
Delhi Technological University on 2026-05-28 <1%
- 2** Student papers
Indian Institute of Technology (ISM), Dhanbad on 2026-05-01 <1%
- 3** Student papers
Delhi Technological University on 2020-05-24 <1%
- 4** Student papers
University of Mosul on 2024-10-23 <1%
- 5** Publication
Hongru Hu, Siddhant Sanghi, Gerald Quon. "Predicting emergent phenotypes fro... <1%
- 6** Student papers
Zhejiang University Center of Modern Educational Technology on 2026-06-05 <1%
- 7** Student papers
Fachhochschule Kärnten Gemeinnützige Privatstiftung on 2025-02-03 <1%
- 8** Student papers
Delhi Technological University on 2026-05-29 <1%
- 9** Internet
arxiv.org <1%
- 10** Publication
Faten S. Alamri, Noor Ayesha, Afia Zafar, Adil Ali Saleem, Amjad R. Khan. "Multim... <1%

11	Student papers	National Institute of Technology, Rourkela on 2026-05-26	<1%
12	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
13	Student papers	Higher Education Commission Pakistan on 2023-09-09	<1%
14	Publication	Song, Lei. "Author Name Disambiguation Using Graph Contrastive Learning", The...	<1%
15	Publication	Abheek Pradhan, Sana Alamgeer, Rakesh Suvvari, Syed Tousiful Haque, Anne H. H...	<1%
16	Student papers	Korea Advanced Institute of Science and Technology on 2025-11-13	<1%
17	Student papers	Faculty of Pharmacy (FOP) on 2026-04-24	<1%
18	Publication	Samah A.F. Manssor, Zhengyun Ren, Rong Huang, Shaoyuan Sun. "Human Activit...	<1%
19	Publication	Farrukh Hafeez, Zeeshan Ahmad Arfeen, Muhammad I. Masud, Abdoalateef Alzhr...	<1%
20	Student papers	Macao Polytechnic Institute on 2025-11-26	<1%
21	Internet	assets-eu.researchsquare.com	<1%
22	Internet	docshare.tips	<1%
23	Student papers	California State University, Sacramento on 2023-11-21	<1%
24	Student papers	Coventry University on 2025-08-05	<1%

25	Student papers	Delhi Technological University on 2026-06-09	<1%
26	Student papers	The Northcap University on 2026-05-15	<1%
27	Student papers	University of Birmingham on 2025-04-17	<1%
28	Student papers	University of Hertfordshire on 2026-05-09	<1%
29	Publication	"Proceedings of 5th International Conference on Frontiers in Computing and Syst...	<1%
30	Student papers	Birla Institute Of Technology & Science - Pilani on 2026-05-03	<1%
31	Student papers	Delhi Technological University on 2026-05-25	<1%
32	Student papers	Delhi Technological University on 2026-06-06	<1%
33	Student papers	Indian Institute of Technology, Madras on 2026-05-12	<1%
34	Publication	Li, Yin. "Fine-Grained Wireless Perception for Human Tracking.", Cornell University	<1%
35	Student papers	Ohio University on 2008-12-15	<1%
36	Publication	Qilong Li, Chongsheng Zhang. "UCR: A unified character-radical dual-supervision ...	<1%
37	Publication	Rytis Maskeliūnas, Audrius Kulikajėvas, Robertas Damaševičius, Julius Griškevičius...	<1%
38	Student papers	University of Lincoln on 2026-05-06	<1%

39	Student papers	University of Surrey on 2024-03-19	<1%
40	Student papers	University of Sussex on 2026-05-06	<1%
41	Student papers	Vietnamese-German University on 2025-07-12	<1%
42	Internet	poster-openaccess.com	<1%
43	Internet	prip.by	<1%
44	Internet	www.cbit.ac.in	<1%
45	Student papers	Coventry University on 2025-04-09	<1%
46	Student papers	Delhi Technological University on 2026-05-06	<1%
47	Student papers	Higher Education Commission Pakistan on 2026-02-03	<1%
48	Publication	Priyanka Gautam, Manjeet Singh. "A multidimensional U-net architecture for Alz...	<1%
49	Student papers	University of Northumbria at Newcastle on 2025-08-28	<1%
50	Student papers	University of Sheffield on 2026-06-01	<1%
51	Student papers	University of St Andrews on 2026-04-03	<1%
52	Publication	Xin Yan, Xiaohang Zhao, Haiyang Chi, Qingwang Wang et al. "FreqMamba: Freque...	<1%

53	Publication	Xu, Yi. "Advancing Autonomous Driving: From Modular to End-to-End", Northeast...	<1%
54	Internet	knowledgecommons.lakeheadu.ca	<1%
55	Internet	link.springer.com	<1%
56	Internet	www.archyworldys.com	<1%
57	Publication	"Computer Vision and Image Processing", Springer Science and Business Media L...	<1%
58	Publication	"Hybrid Artificial Intelligent Systems", Springer Science and Business Media LLC, ...	<1%
59	Student papers	Asia Pacific University College of Technology and Innovation (UCTI) on 2025-05-20	<1%
60	Student papers	Delhi Technological University on 2026-05-20	<1%
61	Publication	Ehab Essa, Islam R. Abdelmaksoud. "Temporal-channel convolution with self-atte...	<1%
62	Student papers	Guru Ghasidas Vishwavidyalaya, Bilaspur on 2026-05-05	<1%
63	Student papers	Higher Education Commission Pakistan on 2025-09-22	<1%
64	Student papers	Imperial College of Science, Technology and Medicine on 2021-09-03	<1%
65	Student papers	Liverpool John Moores University on 2023-06-02	<1%
66	Publication	Mohammad Zeeshan. "Efficient Deep Learning Models for Edge IOT Devices - A Re...	<1%

67	Student papers	Nottingham Trent University on 2026-06-05	<1%
68	Student papers	The University of Manchester on 2026-05-05	<1%
69	Student papers	University of East London on 2025-05-09	<1%
70	Student papers	University of Queensland on 2026-06-06	<1%
71	Publication	Vinay Gautam, Gaganpreet Kaur, Anurag Sinha, Ashish Kumar Verma, Anindya N...	<1%
72	Internet	jurnal.polbeng.ac.id	<1%
73	Internet	oa.upm.es	<1%
74	Internet	par.nsf.gov	<1%
75	Internet	repository.iau.edu.sa	<1%

MULTIMODAL HUMAN ACTIVITY RECOGNITION USING DEEP LEARNING

8
A Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in

SOFTWARE ENGINEERING

by

MANSI GUPTA

(24/SWE/09)

1
Under the Supervision of

Dr. Abhilasha Sharma

Associate Professor, Department of Software Engineering

Delhi Technological University, Delhi, India



Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-110042, India

June, 2026

ACKNOWLEDGEMENTS

**Gurur Brahma Gurur Vishnuh Gurur Devo Maheshwarah I
Guruh Sakshat Param Brahma Tasmai Shri Gurave Namah ||**

I am thankful to the divine for being a pillar of strength, resilience and guidance in this journey. This belief provided clarity and encouragement in the challenges and successes faced throughout the process of this research.

My sincere and heartfelt thanks to my supervisor, Dr. Abhilasha Sharma for her invaluable guidance, constant support and encouragement from the initial stages of this research. Her wise advice, constructive criticism and gentle guidance were instrumental in the direction and quality of this work. Her focus on clarity, rigor and critical thinking has significantly deepened my knowledge and understanding and reinforced my approach to research problems. I am deeply grateful for her constant faith in my capabilities.

I would like to thank Dr. Roshni Singh for her technical inputs, her collaborative approach towards discussions and her constant encouragement throughout the course of my research and M.Tech journey. Her practical insights and thoughtful inputs made a real difference to the development and refinement of key aspects of this work. I am grateful to her for sharing her knowledge and being available to offer guidance when it was needed.

I owe a special thanks to my parents and my elder brother Mayank for their unconditional love, encouragement at every step of this journey and guidance throughout my life. Without their support and motivating presence, it would not have been possible for me to achieve my goals. I have always been able to rely on their trust in me. I am grateful to my friends, colleagues and peers for their companionship, encouragement, and academic camaraderie, which provided balance and positivity throughout this journey. I am grateful to all of them for their support, love and guidance in the completion of this thesis.

Mansi Gupta
24/SWE/09

3

**DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I Mansi Gupta hereby certify that the work which is being presented in the thesis entitled Multimodal Human Activity Recognition Using Deep Learning in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Software Engineering, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own 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.

Candidate's Signature

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

Signature of Supervisor(s)**Signature of External Examiner**

1



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Mansi Gupta** (24/SWE/09) has carried out her research work presented in this thesis entitled **“Multimodal Human Activity Recognition Using Deep Learning”** 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 herself 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.

Place: Delhi

Date:

Dr. Abhilasha Sharma

Associate Professor

Dept. of Software Engineering, DTU

Multimodal Human Activity Recognition Using Deep Learning

Mansi Gupta

ABSTRACT

Human activity recognition has been a subject of intense research, as it is at the cross-roads of several practical areas such as healthcare monitoring, rehabilitation, fitness analysis, smart environments, surveillance, and human-computer interaction. Pose estimation methods are used to extract human joint information from visual and sensor data, which are essential for these systems to represent human motions. While significant strides have been made with deep learning, it is not easy to deploy the technology in unstructured environments. Benchmarking performance does not address all the issues that arise from using human postures, occlusion, environmental conditions and from different users. A successful system must achieve recognition accuracy, be computation efficient, account for the temporal aspects of activity sequences and generalize in a consistent manner across different scenarios. The present study explores the design and deployment of deep learning models to realize robust human activity analysis. Pose estimation architectures from convolutional and transformer-based approaches are assessed based on standard metrics like mean Average Precision, Percentage of Correct Keypoints and Mean Per Joint Position Error on the COCO, MPII and the Human3.6M dataset. More complex spatial relationships are easier to capture with more powerful models, but require more computational power. The performance is significantly influenced by viewpoint variation, occlusion, and the possible bias in the datasets, and this reinforces the fact that the selection of the model should be made based on the requirements of a given deployment, and not just the benchmark.

A skeleton-based model for both spatial structure and temporal motion is proposed to investigate the effect of pose estimation quality on activity recognition. The key points of the skeleton are extracted and converted from several pose estimation backends to spatial and kinematic representations that include the coordinates of the joints and the distances and the angles between them. They have been processed by dual-branch CNN-LSTM architecture which learns spatial patterns and motion dynamics at the same time. MediaPipe Pose sets a new recognition rate of 94% and offers lightweight recognition alternatives like YOLOv11 Pose with efficient inference times of 125.0ms, 8.0fps, and a model size of 3.68MB. The results indicate that accuracy is not the only criterion that should be used in selecting a model, and that the practical limitations of the deployment environment must be taken into account.

In scenarios where visual data is not sufficient, a Multi-Modal Hierarchical Temporal Transformer framework is developed, which integrates the streams of wearable and ambient sensors. An adaptive fusion mechanism dynamically weights the contribution from each sensor to perform the modality specific feature extraction, and the transformer encoder layers with self-attention are used to model long-range temporal dependencies. By using a hierarchical learning strategy, locomotion patterns can be learned concurrently with higher level activities from a shared representation. For Opportunity++, the macro F1 score in Leave-One-Subject-Out cross validation for locomotion and higher-level activity recognition is 0.9290 and 0.8111 respectively, with accuracy of 91.22% and 81.49% respectively. On PAMAP2 and DSADS, accuracy reaches 92.03% and 84.81%. The lightweight version decreases the parameters from 1.16M to 0.009M and the model size from 4.48MB to 0.05MB, enabling deployment on resource-constrained devices. The findings of this research clearly indicate that none of the single modalities or the single model design in the tested scenarios provide reliable activity recognition without any influence in real world conditions. Together, accurate motion representation, well-structured temporal modeling and careful integration of complementary data sources are the building blocks of systems that can support real-world applications, such as in the context of healthcare monitoring, assisted living, smart homes, and context-aware computing.

Keywords: *Human Activity Recognition, Deep Learning, Multimodal Sensor Fusion, CNN-LSTM, Transformer, Skeleton-Based Recognition, Cross-User Generalization, Edge Intelligence, Human Pose Estimation, Wearable Sensors.*

LIST OF PUBLICATIONS

International Conferences

1. Gupta, M. and Sharma, A., 2025. Comparative analysis of skeleton-based pose estimation models for human activity recognition using CNN-LSTM. In *Proceedings of the 7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (MIND-2025)*. (Scopus-Indexed, Accepted and Presented, 13 December 2025)
2. Gupta, M. and Sharma, A., 2026. A comparative analysis of deep learning approaches for human pose estimation: Benchmarks, applications, and future outlook. In *Proceedings of the International Conference on Innovative Computing and Communications (ICICC 2026)*. (Scopus-Indexed, Accepted and Presented, 07 February 2026)

Book Chapter

1. Singh, R., Gupta, M. and Sharma, A., 2026. Innovative applications of visual perception in modern AI systems. In *Advanced Techniques in Computer Vision: Current Research and Emerging Applications*, Chapter 14. Bentham Science Publishers. (Book Chapter, Accepted)

International Journal

1. Gupta, M., Sharma, A., and Singh, R., 2026. LiteHAR-DSC: A Depthwise Separable Convolutional Framework for Edge-Deployable Cross-User Human Activity Recognition on Wearable Sensors. *Advanced Engineering Sciences*, Sichuan University. (Accepted for Publication)
2. Gupta, M., Sharma, A., and Singh, R., 2026. MM-HTT: Cross-user hierarchical temporal learning from wearable and ambient sensor streams. (Under Review)

TABLE OF CONTENTS

Acknowledgements	i
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
List of Publications	vi
Table of Contents	vii
List of Table(s)	xi
List of Figure(s)	xii
List of Symbols, Abbreviations and Nomenclature	xiv
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	4
1.3 Research Objectives	5
1.4 Summary of Contributions	6
1.5 Thesis Organisation	7
2 Literature Review	9
2.1 Evolution of HPE: Traditional vs DL Approaches	9
2.2 CNN-Based and Transformer-Based Approaches for Pose Estimation .	11
2.3 Vision-Based HAR: Skeleton and Appearance Methods	12
2.4 Sensor-Based HAR: Wearable and Ambient Systems	14
2.5 Multimodal Fusion Strategies for HAR	15
2.6 Benchmark Datasets and Evaluation Protocols	16
2.7 Research Gaps	17
2.8 Summary	18
3 Comparative Analysis of Pose Estimation Models	19
3.1 Overview of DL-based Pose Estimation Approaches	19
3.2 Taxonomy of Pose Estimation Architectures	19
3.3 Benchmark Datasets and Evaluation Metrics: Comparative Perspective	21

3.4	Comparative Performance Analysis of Pose Estimation Models	22
3.5	Accuracy and Computational Cost Trade-offs	24
3.6	Application-based Analysis of Pose Estimation Models	26
3.7	Limitations and Future Research Directions	27
3.8	Summary	28
4	Skeleton-Based HAR Using CNN-LSTM	30
4.1	System Architecture	30
4.1.1	Overall Pipeline Description	31
4.1.2	Pose Estimation Backends	32
4.2	Dataset Construction	33
4.2.1	Data Collection and Preprocessing	33
4.2.2	Activity Classes	36
4.3	Keypoint Feature Engineering	37
4.3.1	Normalized Joint Coordinates	37
4.3.2	Inter-Joint Distance Features	38
4.3.3	Joint Angle Representation	38
4.3.4	Temporal Feature Construction	38
4.4	Dual-Branch CNN-LSTM Fusion Architecture	39
4.4.1	Spatial Feature Learning using CNN	39
4.4.2	Engineered Feature Learning Branch	40
4.4.3	Temporal Modeling using LSTM	40
4.4.4	Feature Fusion and Classification	40
4.5	Experimental Setup and Evaluation Metrics	41
4.5.1	Training Configuration	41
4.5.2	Evaluation Metrics	41
4.6	Results and Comparative Analysis	42
4.6.1	Classification Performance	43
4.6.2	Latency and Throughput Analysis	44
4.6.3	Accuracy vs Efficiency Trade-off	45
4.6.4	Training Convergence	45
4.7	Discussion	46
4.8	Summary	46
5	Multimodal Hierarchical Temporal Learning for Cross-User HAR	47
5.1	Problem Formulation and Challenges	47
5.1.1	Inter-Subject Variability	47
5.1.2	Sensor Heterogeneity	48

5.2	Proposed Multimodal Learning Framework	49
5.2.1	Multimodal Inputs: Wearable and Ambient Sensors	51
5.2.2	Modality-Specific Convolutional Encoders	51
5.2.3	Learnable Attention-Based Fusion	53
5.2.4	Transformer Encoder for Temporal Modeling	55
5.3	Hierarchical Multi-Task Learning	56
5.3.1	Locomotion-Level Activities	56
5.3.2	High-Level Daily Activities	56
5.4	Experimental Setup	57
5.4.1	Datasets: Opportunity++, PAMAP2, DSADS	57
5.4.2	Training Configuration and Implementation Details	59
5.4.3	LOSO Evaluation Protocol	61
5.5	Lightweight Model for Edge Deployment	61
5.6	Results and Analysis	63
5.6.1	Overall Performance	63
5.6.2	Cross-User Generalization	65
5.6.3	Effect of Multimodal Fusion and Temporal Modeling	65
5.7	Discussion	65
5.8	Summary	66
6	Results and Comparative Discussion	68
6.1	Comparative Analysis Across the Three Studies	68
6.2	Vision-based and Sensor-based Activity Recognition: Trade-offs	69
6.3	Accuracy, Latency, and Deployment Feasibility	71
6.4	Pipeline from Pose Estimation to Sensor-based Activity Recognition	73
6.5	Limitations	73
6.6	Summary	75
7	Conclusion, Future Work, and Social Impact	76
7.1	Conclusion	76
7.2	Future Work	77
7.3	Social Impact and Ethical Considerations	77
7.3.1	Healthcare and Rehabilitation	77
7.3.2	Elderly Care and Assistive Systems	78
7.3.3	Fitness and HCI	78
7.3.4	Intelligent Surveillance and Workplace Safety	78
7.3.5	Privacy and Ethical Concerns	78

References	79
List of Publications and their Proofs	86
Plagiarism Report	100
Curriculum Vitae	101

LIST OF TABLE(S)

2.1	Benchmark datasets used for HAR and HPE	17
2.2	Positioning of existing works with respect to identified research gaps .	18
3.1	Evaluation metrics used in HPE research	22
3.2	Temporal progression of representative HPE models on the COCO validation set	23
4.1	Pose estimation backends used for keypoint extraction	32
4.2	Preprocessing specifications for each pose estimation backend	33
4.3	Summary of the constructed HAR dataset	34
4.4	Balanced data distribution across activity classes	35
4.5	Keypoint distribution across pose estimation backends	36
4.6	Input dimensions for the CNN-LSTM fusion architecture	39
4.7	Classification performance of the CNN-LSTM fusion architecture across pose estimation backends	43
4.8	Computational efficiency of CNN-LSTM models evaluated on CPU .	45
5.1	Summary of datasets used for MM-HTT evaluation	58
5.2	Efficiency comparison between the full MM-HTT framework and the lightweight model	62
5.3	Sensor modality ablation results on the Opportunity++ dataset	63
5.4	Sensor ablation results on DSADS and PAMAP2 datasets	64
5.5	LOSO cross-user evaluation results	65
6.1	Comparison across the three approaches based on input representation, model design, and performance characteristics	69
6.2	Deployment characteristics across the analyzed approaches	72
6.3	Summary of limitations across the three studies	75

LIST OF FIGURE(S)

1.1	Key application domains of HAR	2
1.2	General pipeline for skeleton-based HAR	3
1.3	Mapping of identified research challenges to contributions of this research work	4
2.1	Evolution of HPE from classical methods to transformer-based approaches	10
2.2	Comparison between CNN-based and transformer-based pose estimation architectures	12
2.3	Comparison between appearance-based (RGB) and skeleton-based HAR approaches	13
3.1	Taxonomy of HPE architectures	20
3.2	Accuracy-complexity relationship of representative HPE models on the COCO validation set (mAP vs. relative model complexity)	24
3.3	Trade-off among accuracy, computational cost, and deployment efficiency in HPE models	25
4.1	Pipeline of the proposed skeleton-based HAR architecture	31
4.2	Representative RGB frames and corresponding skeletal keypoint representations for the four activity classes	34
4.3	Illustration of spatial features derived from skeletal keypoints, including normalized joint coordinates, inter-joint distances, and joint angles	37
4.4	Dual-branch CNN-LSTM fusion architecture for skeleton-based HAR	40
4.5	Confusion matrices of the four pose estimation backends on the four-activity test dataset	44
4.6	Training and validation accuracy trends of the CNN-LSTM fusion model across all pose estimation backends	46
5.1	Complementary characteristics of wearable and ambient sensor modalities motivating multimodal fusion	48
5.2	Architecture of the proposed multimodal hierarchical temporal learning framework	50
5.3	Modality-specific convolutional encoders for wearable and ambient sensor streams with temporal feature extraction	52
5.4	Attention-based scalar fusion of wearable and ambient features using learnable modality weights and fusion projection	54

5.5	Confusion matrix and one-vs-rest ROC curves for the fusion configuration on the Opportunity++ dataset	64
5.6	Confusion matrices and ROC curves for DSADS and PAMAP2 datasets under the fusion configuration	66
6.1	Comparison between vision-based and sensor-based HAR approaches	70
6.2	Comparison of recognition performance and model complexity across the analyzed approaches	71
6.3	Progression from HPE to multimodal HAR	73

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

Symbols	Description
α	Attention weight for modality fusion
θ	Trainable model parameters
η	Learning rate of the optimizer
ϵ	Small constant for numerical stability / label smoothing
$\hat{x}_i$	Normalized x-coordinate of joint i
$\hat{y}_i$	Normalized y-coordinate of joint i
θ_B	Joint angle at point B
d_{ij}^{norm}	Normalized Euclidean distance between joints i and j
F_{frame}	Frame-level feature vector
$F_{sequence}$	Temporal sequence of stacked frame features
X	Input multivariate sensor sequence
X_w	Wearable sensor feature stream
X_a	Ambient sensor feature stream
H_w	Wearable modality feature representation
H_a	Ambient modality feature representation
H_f	Fused multimodal feature representation
Q	Query matrix in transformer attention
K	Key matrix in transformer attention
V	Value matrix in transformer attention
z	Shared pooled feature representation
$\hat{y}$	Predicted activity label
Abbreviations	Full Form
HAR	Human Activity Recognition
HPE	Human Pose Estimation
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
DL	Deep Learning
RGB	Red Green Blue
fps	Frames Per Second
PAF	Part Affinity Field
GCN	Graph Convolutional Network
CNN-LSTM	Convolutional Neural Network with Long Short-Term Memory

ST-GCN	Spatial Temporal Graph Convolutional Network
LOSO	Leave-One-Subject-Out
MM-HTT	Multi-Modal Hierarchical Temporal Transformer
IoT	Internet of Things
SOTA	State-of-the-Art
AR	Augmented Reality
VR	Virtual Reality
HCI	Human-Computer Interaction
mAP	Mean Average Precision
OKS	Object Keypoint Similarity
MPJPE	Mean Per Joint Position Error
PCK	Percentage of Correct Keypoints
PCPm	Percentage of Correctly Predicted Body Parts
AUC	Area Under Curve
NME	Normalized Mean Error
ROC	Receiver Operating Characteristic
ReLU	Rectified Linear Unit
FFN	Feed Forward Network
Conv1D	One-Dimensional Convolution
MoCap	Motion Capture
DNN	Deep Neural Network
GPU	Graphics Processing Unit
CPU	Central Processing Unit
COCO	Common Objects in Context
CV	Computer Vision
LiDAR	Light Detection and Ranging
DCAM-Net	Deformable Convolution and Attention Mechanism Network
DSADS	Daily and Sports Activities Dataset

Nomenclature

Meaning

x_i, y_i	Spatial coordinates of body joint i
W, H	Frame width and frame height
d_{ref}	Reference distance used for normalization
A, B, C	Adjacent joints used for angle computation
h_t	Hidden state of LSTM at time step t
Z_t	Concatenated spatial and kinematic features at time step t
o_{loc}	Logits for locomotion prediction

o_{hl}	Logits for high-level activity prediction
L	Total training loss
L_{loc}	Locomotion classification loss
L_{hl}	High-level activity classification loss
w_{loc}	Weight assigned to locomotion loss
w_{hl}	Weight assigned to high-level activity loss
T	Temporal window length
S	Sliding window stride
N	Number of segmented windows
F	Number of input sensor features
D	Embedding dimension of feature representation

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

One of the more challenging open problems in artificial intelligence and computer vision (CV) is the understanding of human movement from sensor data. Human Activity Recognition (HAR) is the field where researchers are trying to enable machines to recognize human activity: walking, sitting, falling, or even subtle daily activities like eating with data collected via camera, wearable sensor, or ambient devices placed in the environment [1–3]. HAR has come a long way in the past 20 years. Now it has common applications outside of the laboratory health monitoring, assisted living, smart home systems, surveillance and HCI [4–6] and is being done using commercially available hardware and appropriate algorithms rather than custom-designed research rigs.

In the field of healthcare and rehabilitation, automatic activity recognition allows for the ability to continuously monitor a patient’s physical state, monitor recovery progress, and monitor if a patient is following the exercise programs [7,8]. HAR systems can provide objective and data-driven information about the movement of the joints and the postural behavior, eliminating the need for frequent face-to-face assessments and enabling remote monitoring of patients who are elderly or have reduced mobility [9, 10]. HAR is also being used in intelligent systems for fall detection, behaviour monitoring of elderly people and automatic action for safety and comfort in smart homes and assisted living environments [2, 11].

The application of HAR in combination with surveillance holds promise for future implementation of behavior analytics in public settings and security-sensitive areas, where automated posture and crowd recognition can complement traditional surveillance methods in a meaningful way [12]. In the field of HCI, gesture and activity recognition are becoming common methods of human interaction with immersive interfaces, augmented reality (AR) environments, and accessibility applications for persons with physical disabilities [13].

There are two key research approaches that are prevalent in the HAR field. The first is the *vision-based* approach that uses vision cues from images and video streams to determine activity. The second approach is the *sensor-based* one, where the signal from sensors attached in the environment is analyzed. Vision-based techniques, particularly vision-based deep learning (DL) and human pose estimation (HPE) methods, offer a wealth of spatial information of the human body. A subset of these methods are skeleton-based methods which have been found to be more successful because they model the geometry of the joint configurations rather than the appearance, and are thus more robust to changes in lighting, clothing and viewpoint, as mentioned in [14, 15]. There are some advantages, however, to sensor-based techniques. Less privacy intrusive, relatively low energy consumption and can operate in uncontrolled outdoor environments where it is impractical to use cameras. They do, however, have their own challenges, particularly when it comes to the inter-subject variations and noise in real-world data acquisition scenarios [7, 16]. Both have difficulty in creating efficient, generalizable, robust recognition systems that can work outside of the laboratory and are not time consuming. As shown in Fig. 1.1, HAR could have interesting applications in various areas such as smart environments, surveillance, human-computer interaction (HCI) and healthcare.

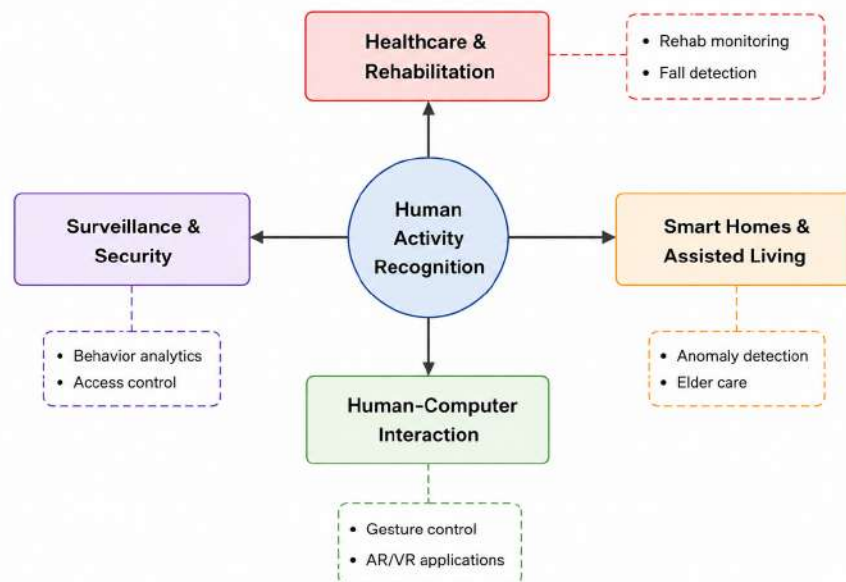


Figure 1.1: Key application domains of HAR

Increasingly complex application domains have led to a rethinking of methodological decisions by researchers. Previous activity recognition systems were based on hand-

crafted features extracted from silhouettes, contours and optical flow. These performed fairly well in controlled lighting, but they failed to perform well when the lighting conditions changed, when objects occluded subjects, or when the background became cluttered [1]. The introduction of DL brought a new way of creating representations, which is automated, instead of the traditional way of manual engineering. The feature extraction process is now directly performed by convolutional neural networks (CNNs), recurrent neural networks (RNNs) and more recently, transformers that learn features from raw sensor data. In both vision and sensor-based approaches, the performance of recognizers has improved greatly [12, 15, 17–19].

Today's HAR pipeline typically consists of four distinct steps pose estimation, feature extraction, temporal modeling, and activity classification as shown in Fig. 1.2. The stages are tightly coupled. So, the quality of decisions made at the beginning of the pipeline influences the system's final recognition.

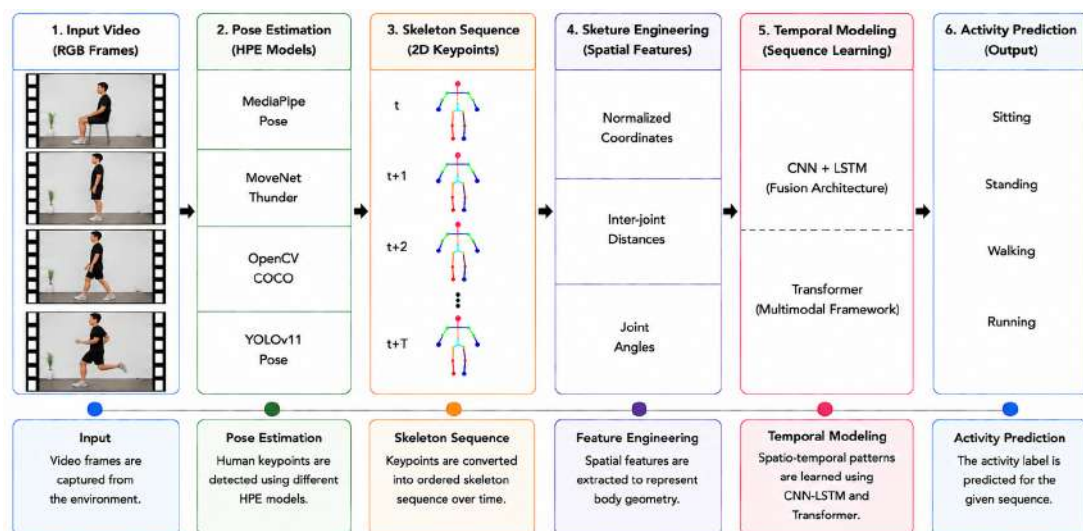


Figure 1.2: General pipeline for skeleton-based HAR

Despite these advances, several fundamental challenges remain unresolved. The tension between *recognition accuracy* and *real-time computational efficiency* continues to limit deployment on resource-constrained edge devices [13, 20, 21]. *Cross-user generalization* the ability of a model trained on one set of individuals to recognize activities performed by unseen users remains one of the most difficult problems in sensor-based HAR, owing to the substantial variability in how different individuals perform the same activity [4, 22, 23]. In vision-based HAR, *occlusion and viewpoint variation* introduce significant noise into skeletal keypoint estimates, while the lack

of standardized benchmarks makes comparative evaluation across methods difficult [24–27]. These challenges collectively motivate the research presented in this work.

1.2 Problem Statement

While significant progress has been made in both HPE and HAR, three interconnected problems continue to limit the applicability of existing methods in real-world settings. Figure 1.3 illustrates the relationship between these challenges and the contributions of this research work.

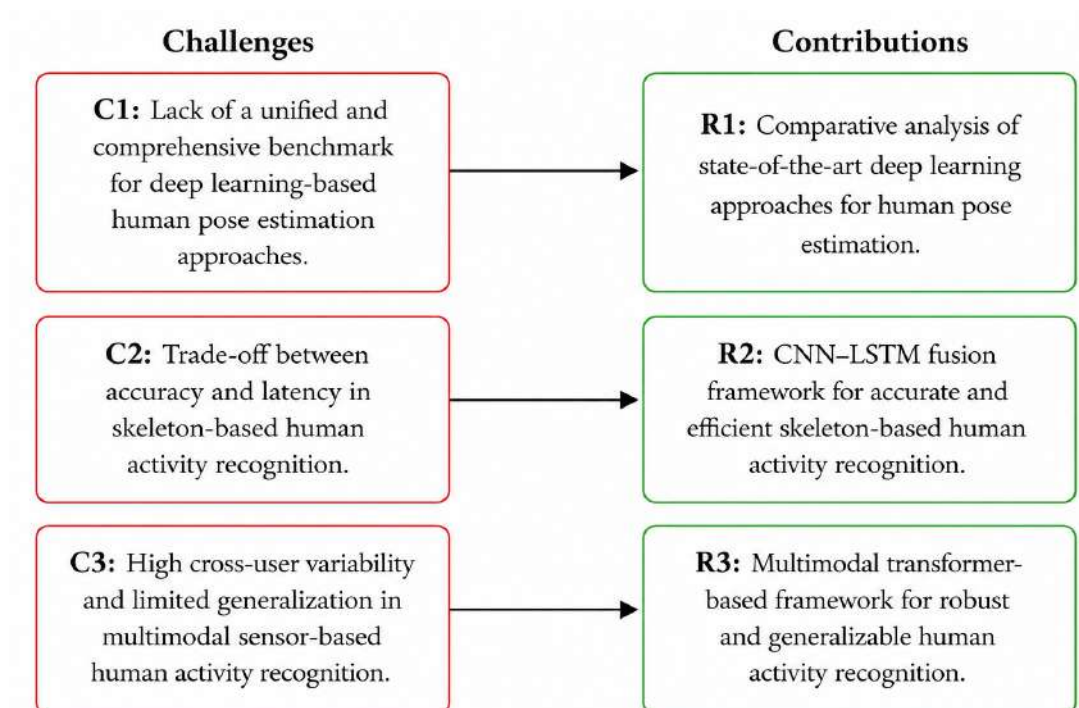


Figure 1.3: Mapping of identified research challenges to contributions of this research work

Problem 1: Fragmented evaluation of DL approaches to HPE. A large number of DL models for HPE have been proposed in the literature, ranging from early convolutional approaches such as OpenPose [24] and HRNet [28] to transformer-based systems including ViTPose [29] and PoseFormerV2 [25]. However, these models are evaluated across heterogeneous benchmark datasets using differing evaluation metrics, making systematic cross-model comparison difficult. There is a need for a consolidated comparative analysis that contextualizes model performance alongside computational cost and deployment feasibility, particularly for practitioners seeking to select appropriate architectures for specific real-world applications [26, 30].

Problem 2: Lack of standardized comparison of skeleton-based HAR across pose estimation backends. The success of skeleton-based HAR systems heavily relies on the accuracy of the employed pose estimation model that is used for keypoints extraction. Pose backends like MediaPipe Pose [13], MoveNet [20], OpenCV COCO [24] and YOLOv11 Pose [31] vary greatly in terms of the number of keypoints they produce, occlusion sensitivity, speed, and memory usage. Although these are important distinctions, the majority of studies only assess one pose backend and assume it is a fixed value and not something to consider. This is a real challenge in understanding the practical consequences of selecting one backend over the other. The current problem with the field is the absence of a shared comparison platform that lets multiple backends undergo the identical experiments and allows them to be fairly and systematically compared to each other [15, 32, 33].

Problem 3: Cross-user generalization in multimodal sensor-based HAR. When applied to a new set of people, HAR systems that are trained using a fixed set of subjects often fail to achieve the same performance. This is mainly because people display significant variation in their movement, the placement of sensors on the body and the way they personally perform the same activity, as described in [5, 23]. Some work has been done in this area so far with domain adaptation and subject independent cross validation methods. But few frameworks are able to resolve the issue of multimodal fusion of wearable sensors and ambient sensors, model the temporal dependency between them and the hierarchical structure of human activities in a single end-to-end architecture [19, 34, 35]. But, aside from this, there's a deployment challenge which is often overlooked. Many real-world HAR applications are executed on edge devices with limited resources like smart phones or embedded sensors, with a strict computational budget. Meanwhile, models need to be designed to achieve good recognition accuracy in dramatically reduced parameter counts, which directly contradicts the complexity required to address the problems mentioned above [21, 36].

1.3 Research Objectives

This research work aims at achieving the following objectives to address the challenges identified above:

RO1. To conduct a comparative analysis of DL-based HPE approaches, starting from the convolutional network toward the transformer-based system, different available benchmark datasets, and different evaluation metrics, and discuss the pros and cons of the accuracy, computational complexity, and the suitability for

real-world applications.

RO2. To design and evaluate a comparative approach for skeleton-based HAR that integrates four popular pose estimation back-ends, namely MediaPipe Pose, MoveNet Thunder, OpenCV COCO, and YOLOv11 Pose into a common dual-branch CNN-LSTM fusion network. The idea is that all of the backends can be tested under the same experimental setup, and then compared in terms of accuracy, latency, throughput and model size.

RO3. To propose and evaluate a multimodal temporal learning framework for cross-user HAR, by using feature-level fusion of multiple sensor modalities and capturing temporal dependency with DL architectures. The goal also includes systematic evaluation in the cross-user configurations of all publicly available datasets, while still considering the computational efficiency of the resulting models, with the aim to ensure that they remain usable on resource-constrained devices.

1.4 Summary of Contributions

This research work makes three primary contributions, each corresponding to a major research work. The contributions are summarized below:

Contribution 1: Comparative analysis of DL approaches for HPE. A comprehensive comparative review of DL-based HPE is presented, tracing the methodological evolution from classical convolutional architectures to transformer-based models. The review provides structured analysis of benchmark datasets (COCO [37], MPII [38], Human3.6M [39]), evaluation metrics (mAP, PCK, MPJPE), and quantitative model performance across CNN-based, lightweight, hybrid, and transformer-based approaches. The survey covers four application domains healthcare, rehabilitation, surveillance, robotics, and what each actually demands from a deployed system. It also identifies the areas that the field is still struggling with: occlusion handling, cross domain generalization, and dataset diversity.

Contribution 2: Skeleton-based HAR using a CNN-LSTM fusion framework. A CNN-LSTM fusion architecture is designed and tested for the implementation of skeleton based HAR with four pose estimation backends MediaPipe Pose [13], MoveNet Thunder [20], OpenCV COCO [24], and YOLOv11 Pose [31] within a single pipeline. The dataset was balanced with four activities (sitting, standing, walking and running).

walking, and running) was constructed, and keypoints were transformed into two feature vectors spatial and kinematic are utilized for the representation of normalized coordinates, inter-joint Euclidean distances, pairwise and joint angular relationships. Under the same experimental conditions, the best accuracy is 94% achieved by MediaPipe, with YOLOv11 Pose having the lowest latency of 125.0 ms at 8.0 fps. These results establish a reproducible benchmark for pose backend selection in real-time HAR systems.

Contribution 3: Multimodal Hierarchical Temporal Transformer (MM-HTT) for cross-user HAR. The MM-HTT framework is proposed for cross-user HAR using synchronized wearable and ambient sensor streams. The framework employs modality-specific one-dimensional convolutional encoders, a learnable scalar attention-based modality weighting mechanism, and stacked transformer encoder layers for long-range temporal dependency modeling within fixed-length sensor windows. A hierarchical multi-task learning strategy enables simultaneous prediction of locomotion primitives and high-level daily activities from a shared temporal representation. Evaluated using Leave-One-Subject-Out (LOSO) cross-validation on three public datasets (Opportunity++ [40], PAMAP2 [41], DSADS [42]), the framework achieves a macro F1-score of 0.9290 for locomotion and 0.8111 for high-level activity recognition on Opportunity++, and up to 92.03% accuracy on PAMAP2. A lightweight variant reduces parameters from 1.16M to 0.009M and model size from 4.48 MB to 0.05 MB.

6

1.5 Thesis Organisation

The remainder of this research work is organized as follows.

Chapter 2 presents a comprehensive literature review covering the evolution of HPE and HAR, spanning traditional feature-based methods, CNN-based architectures, transformer models, and multimodal sensor fusion strategies. Across-field benchmark datasets and evaluation protocols are reviewed and key gaps in research identified.

Chapter 3 presents the first research contribution: a systematic comparative analysis of DL-based HPE approaches. The chapter reviews architectural taxonomies, benchmark datasets, evaluation metrics, and performance trade-offs across model families, and maps findings to application-specific deployment considerations.

Chapter 4 presents the second research contribution: a comparative evaluation of skeleton-based HAR using four pose estimation backends integrated within a dual-branch CNN-LSTM fusion architecture. Dataset construction, keypoint feature engi-

neering, architecture design, and experimental results are described in detail.

Chapter 5 presents the third research contribution: the MM-HTT framework for cross-user multimodal HAR. The chapter covers problem formulation, framework architecture, experimental setup, LOSO evaluation results, sensor modality ablation studies, and the lightweight deployment variant.

Chapter 6 provides a cross-work analytical discussion, comparing the three contributions along the dimensions of accuracy, computational efficiency, deployment feasibility, and generalization. A unified conceptual pipeline connecting HPE to skeleton-based HAR to sensor-based HAR is presented, and limitations common to all three studies are examined.

Chapter 7 concludes the research work with a summary of contributions, key findings, social impact and ethical considerations, limitations, and directions for future research.

6

25

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the body of research that forms the foundation of the present study. The review is organized to reflect the progression of methodologies from HPE to skeleton-based activity recognition and further to multimodal sensor-based recognition. The chapter begins with traditional and DL-based approaches for pose estimation, followed by advancements in convolutional and transformer-based models. It then examines vision-based HAR using both appearance and skeleton representations and extends to sensor-based approaches, including wearable and ambient systems. Multimodal fusion strategies are subsequently discussed, followed by an overview of benchmark datasets and evaluation protocols [2, 23, 27]. The chapter concludes by identifying the key research gaps that motivate the proposed work and experimental investigations presented in the subsequent chapters.

2.1 Evolution of HPE: Traditional vs DL Approaches

HPE is the task of localizing anatomical keypoints such as joints of the shoulders, elbows, wrists, hips, knees, and ankles from images or video frames, and representing their spatial relationships as a skeletal model [29, 43]. The evolution of HPE has followed the broader trajectory of CV research, transitioning from geometry-driven, handcrafted approaches to end-to-end DL frameworks capable of achieving real-time performance across diverse and unconstrained environments. The classical methods for HPE had parts-based and probabilistic graphical models as their basis. The human body was usually modeled as a set of rigid components linked by springy constraints, and the configuration of the joints were optimized by dynamic programming over a graph of the body structure built in a tree-like manner [38, 44]. These techniques were valid and fairly effective in the lab. However, their performance breaks down when they are exposed to occlusion, changes of clothing, cluttered backgrounds or changes of viewpoint. They relied a lot on handcrafted features like histogram of oriented gradients (HOG) and scale-invariant feature transform (SIFT), and were very difficult to adapt to new environments without extensive re-engineering work [45].

DL solved these problems in a basic level, by enabling the model to learn hierarchical features directly from the raw image pixels, without manually designed representations of them. Together with the introduction of DeepPose [17,46], one of the earliest CNN-based joint coordinate regression approaches was introduced. It was subsequently found that heatmap-based detection, which predicts the spatial probability distribution of the keypoint location instead of merely a single coordinate, was much more effective in capturing spatial uncertainty. The Part Affinity Field (PAF) [24] was built upon, which opened the door to multi-person pose estimation by providing a vector field representation of limb orientations and allowing the efficient bottom-up grouping of keypoints into individual skeletons. Figure 2.1 summarizes how HPE methods have evolved through these stages.

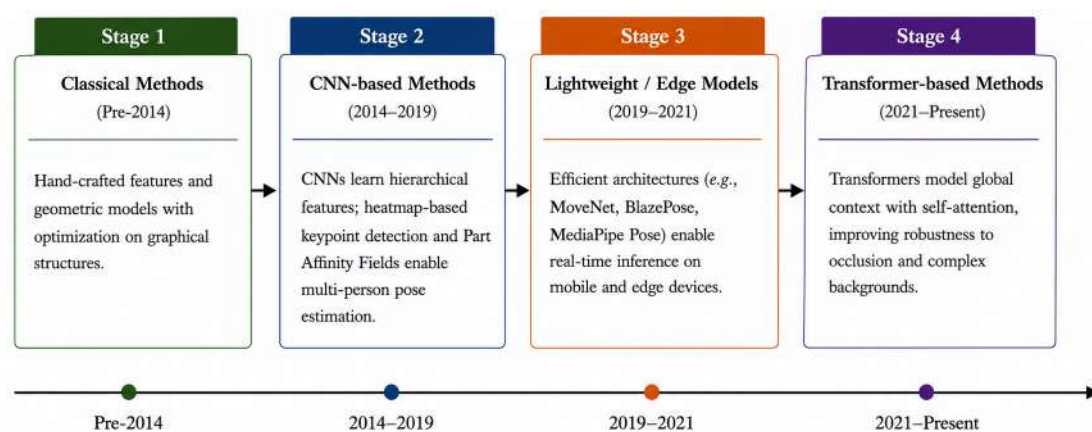


Figure 2.1: Evolution of HPE from classical methods to transformer-based approaches

With the move to DL, there were also positive developments in the multi-person estimation area. There have been two general approaches, with their own practical variations. The top-down approaches like Mask R-CNN train a single-person pose estimator in each region marked by a bounding box detector to detect people [47,48]. This generally results in high accuracy, although the number of people in the scene increases the computational cost, potentially being a significant challenge in crowded settings. Bottom-up approaches like OpenPose able to detect all keypoints of an image simultaneously, and then combine them into person instances [24]. The benefit is that the inference time is remains relatively stable in case of different crowd densities. Both approaches have been continuously improved over time, with the incorporation of high-resolution feature representations, as well as more well-designed loss functions that are better at supervising the learning process [28].

2.2 CNN-Based and Transformer-Based Approaches for Pose Estimation

The High-Resolution Network (HRNet) [28] represented a significant architectural advance over earlier hourglass-style networks by maintaining high-resolution feature representations throughout the entire network rather than recovering resolution through upsampling after feature compression. By connecting parallel multi-resolution streams with repeated multi-scale feature fusion, HRNet achieved state-of-the-art (SOTA) accuracy on the COCO and MPII benchmarks at the time of its introduction [28,37,38]. Similarly, the Multi-Stage Pose Network (MSPN) improved upon earlier designs through multi-stage refinement with cross-stage feature aggregation, demonstrating that iterative keypoint refinement is a robust strategy for handling ambiguous or occluded poses [29].

Two different types of lightweight models have been developed in order to address the large computational power requirement of high accuracy networks. Google's MoveNet [20] is available in two models Lightning for speed and Thunder for accuracy and both utilize depthwise separable convolutional operations with efficient backbones that are optimized for mobile inference. The MediaPipe BlazePose model [13] employs a two-step pipeline. First, a lightweight detecting model detects the presence of a person, and then a regression model provides 33 predicted body keypoint locations. These models provide enough speed for real-time use on mobile devices while providing enough accuracy for use in rehabilitation and fitness applications [9, 10].

Pose estimation with transformers provided a new way of thinking about the problem of pose estimation using our previous methods as a guide. TokenPose [29] uses token embeddings that represent body joints as learnable embedding vectors. Each embedding is then processed together with the embedding vector representations of the various image patch tokens through the encoders in the transformer and as a way for the model to learn how to connect bones. ViTPose [29] demonstrated that vision transformer based backbones already existing in large models pretrained on ImageNet can transfer to pose estimation, showing strong performance on COCO [37] datasets. In 3D pose estimation, PoseFormer [49] applies temporal and spatial transformers to both video sequence of the 2D keypoints to estimate the 3D pose from the 2D keypoints PoseFormerV2 [25] further optimizes the model to reduce the compute time of the temporal modeling by using the frequency domain. The trade-off with using a transformer architecture versus a CNN based architecture is cost. Transformers generally require more computing and memory than either of the other two conditions mentioned. Thus there is a limitation for pose estimation using transformers in resource constrained environments [21,26]. The below Figure 2.2 shows a structural comparison

between each of these models: convolutional output feature extraction versus global attention based feature extraction.

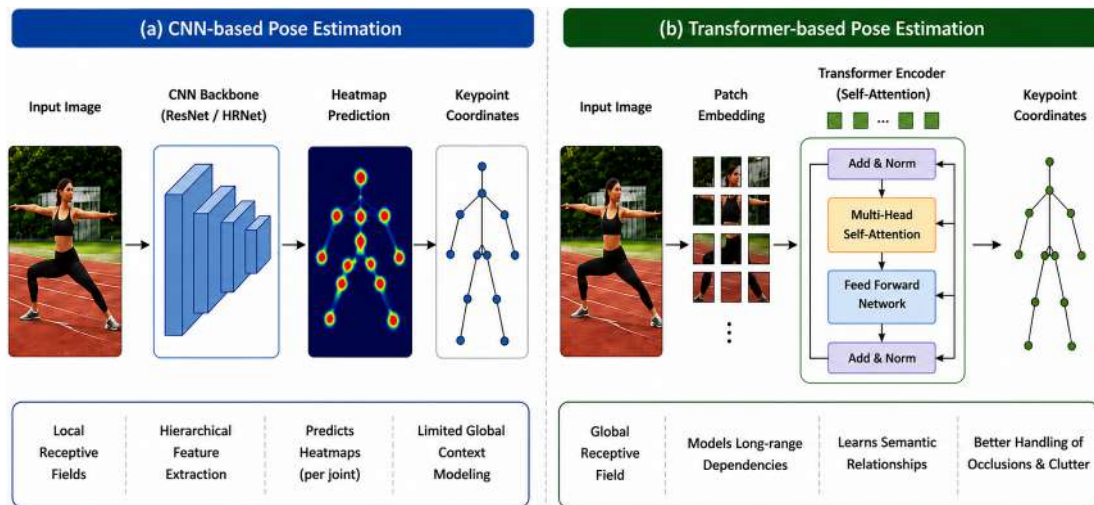


Figure 2.2: Comparison between CNN-based and transformer-based pose estimation architectures

2.3 Vision-Based HAR: Skeleton and Appearance Methods

Vision-based HAR splits broadly into appearance-based and skeleton-based approaches [1, 14]. Appearance-based approaches are applied directly to the pixels optical flow, dense trajectories, CNN-extracted spatial features to describe motion [17, 50]. RGB-D methods have also demonstrated the effectiveness of dynamic image representations in capturing the spatiotemporal patterns [51]. However, pixel-level features are fragile; variation in illumination, background noise, variation in clothes, and change in view-point all have a negative impact on accuracy, making it hard to generalize to real-world conditions. Skeleton-based methods operate differently by taking the sequences of the coordinates of the joints of the moving object from the pose estimator. It is compact, viewpoint-independent, and removes visual clutter that can cause appearance-based methods [14, 43] to fail. This also allows access to geometrically significant features, such as inter-joint distance, joint angles, and orientations of the limbs. The first approaches based on the skeleton extracted these features manually and passed them to SVMs or random forests, with the extent to which these features could be represented being limited by the expressiveness of the hand-designed descriptors.

DL skeleton-based HAR significantly improved. The widespread trend was to use a CNN-LSTM architecture, which uses convolutional layers to capture the local spatial structure in frames or short windows and LSTM layers or bidirectional LSTM layers

to capture the temporal evolution of motion [50, 52]. InnoHAR employed hierarchical feature learning on the stream of wearable and inertial sensors, with the hierarchical layers contributing to the feature learning for complex activities that single-scale features lacked of [53]. In ConvST-LSTM-Net [15] it was argued that the dependencies should be modeled jointly by stacking (alternating) CNN and LSTM layers instead of modeling the spatial and temporal dependencies separately. It introduced spatio-temporal attention by using the STAD-ConvBi-LSTM [12] which allows the model to assign weights to the joints and time steps based on their usefulness for classification.

Graph Convolutional Networks (GCNs) treat the skeleton as a graph joints as nodes, bones as edges and learn from that structure directly [33]. Part-aware ST-GCNs [33] extended this by grouping joints into body parts and using multi-stream architectures to handle occlusion. DCAM-Net [54] added contextual attention across spatial and temporal dimensions, targeting activities where the signal is in small joint movements rather than gross limb motion. One consistent gap across this work: most studies evaluate a single pose estimation backend and never ask how much the choice of pose extractor actually affects downstream HAR performance [30, 32]. That question motivates this study. Figure 2.3 shows the structural difference between appearance-based and skeleton-based approaches.

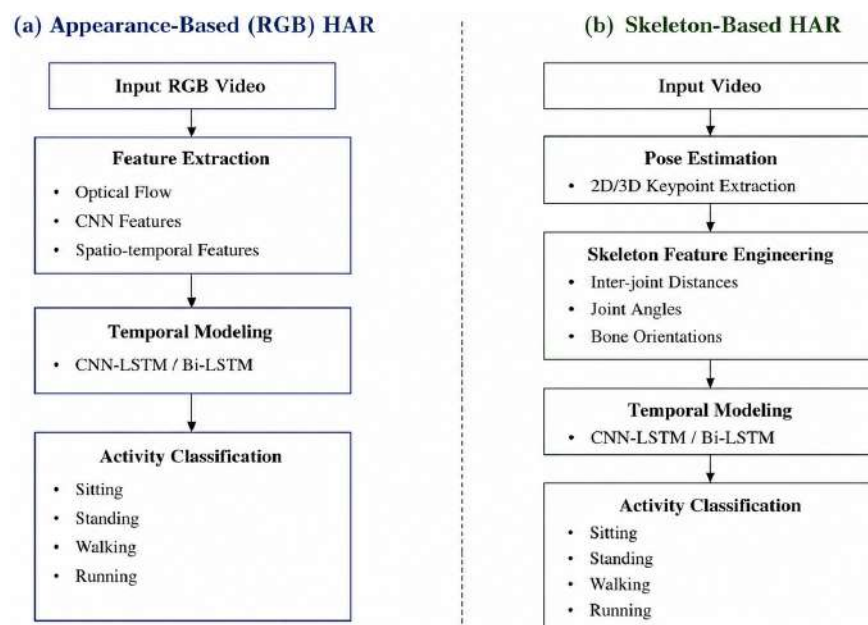


Figure 2.3: Comparison between appearance-based (RGB) and skeleton-based HAR approaches

2.4 Sensor-Based HAR: Wearable and Ambient Systems

Sensor-based HAR avoids the privacy problems of cameras and deploys more easily in everyday settings which is why it has seen wide adoption in healthcare monitoring, smart homes, and assisted living [2, 5, 7]. Two sensing modalities are in common use: wearables on the body and ambient sensors in the environment. Recent work increasingly combines both, since each covers gaps the other misses [55].

Wearable sensors, typically IMUs combining accelerometers, gyroscopes, and magnetometers, sit on the wrist, ankle, chest, or waist and measure body kinematics directly. The data is detailed enough to reliably distinguish locomotive activities and periodic physical behaviors like gait cycles or repetitive exercise movements [7, 16, 56]. Early wearable HAR systems relied on handcrafted statistical features extracted from fixed-length signal windows including mean, variance, root mean square, and spectral coefficients paired with traditional classifiers. DL approaches replaced handcrafted features with end-to-end learned representations: CNN-based models capture local temporal patterns from multivariate inertial signals [18], while LSTM-based architectures model longer sequential dependencies [15, 52]. Deep convolutional architectures such as iSPLInception further demonstrated the effectiveness of inception-residual learning for wearable sensor-based HAR [57]. However, wearable sensors alone are insufficient to recognize activities that are defined by their environmental context. For instance, distinguishing preparing coffee and having breakfast requires contextual cues that body motion alone cannot provide.

Ambient sensors including object interaction sensors, reed switches, RFID tags, microphones, and Wi-Fi or radar-based motion detectors deployed in instrumented environments provide complementary contextual information about a person's interaction with their surroundings [40, 58, 59]. Ambient sensors excel at capturing high-level activity context but typically lack the temporal resolution and motion detail available from inertial wearable devices. In the Opportunity++ dataset [40, 60], for instance, ambient sensors deployed in a kitchen environment recorded object manipulation events that substantially improved the recognition of high-level daily activities when combined with wearable inertial data.

The complementary nature of wearable and ambient sensing modalities motivates their joint use in multimodal HAR frameworks, as explored in recent multimodal approaches [34, 61, 62]. Real-time multimodal activity recognition frameworks have further demonstrated the practical advantages of integrating heterogeneous sensing

modalities within unified DL architectures [63]. Recent multimodal HAR systems have also explored heterogeneous sensing combinations including LiDAR-based environmental sensing for robust activity recognition [64].

2.5 Multimodal Fusion Strategies for HAR

The integration of complementary sensing modalities in HAR has been approached through three principal fusion strategies: *early fusion*, *late fusion*, and *intermediate fusion* [23, 62, 65]. **Early fusion** concatenates features from multiple modalities at the input or feature extraction stage before passing them to a shared model. While simple to implement, this approach may not adequately account for the different sampling characteristics and temporal structures of heterogeneous sensor types [62]. **Late fusion** trains independent models for each modality and combines their predictions through score aggregation or voting. Although this decoupling is robust to modality-specific noise, it does not exploit inter-modal complementarity during representation learning [65]. **Intermediate fusion**, which learns to combine modality-specific representations at an intermediate encoding stage, has been shown to balance the advantages of both strategies [34]. Adaptive or attention-based intermediate fusion mechanisms go further by dynamically weighting modality contributions based on learned importance scores, enabling the model to prioritize the more informative modality for a given activity and sensor context [35]. Recent multimodal HAR systems have also explored heterogeneous sensing combinations including LiDAR-based environmental sensing for robust activity recognition [64].

Transformer-based temporal modeling has recently been applied to sensor-based HAR to address the limitations of CNN and LSTM architectures in capturing long-range temporal dependencies across activity sequences [19, 35, 66]. Self-attention mechanisms enable the direct modeling of relationships between any two time steps within a window, irrespective of their distance, which is particularly valuable for activities that involve extended preparatory phases or complex transition dynamics. The MM-HTT framework combines attention-based intermediate fusion with transformer-based temporal encoding and hierarchical multi-task learning, representing one of the few end-to-end architectures to address all three dimensions simultaneously [35]. Most prior multimodal HAR studies address only one or two of the following dimensions: modality fusion, temporal modeling, hierarchical activity structure, and cross-user generalization [5, 23]. The MM-HTT framework is designed to address all four within a single architecture.

2.6 Benchmark Datasets and Evaluation Protocols

The choice of benchmark dataset and evaluation protocol has a substantial influence on the reported performance of HAR and HPE systems, and comparisons across methods are only meaningful when the methodology is consistent [5, 29]. This section reviews the major benchmark datasets and evaluation protocols relevant to the research work presented in this study.

HPE datasets. The COCO Keypoints dataset [37] contains over 150,000 annotated person instances with 17 keypoints each, covering multi-person, partially occluded, and diverse viewpoint scenarios. The MPII Human Pose dataset [38] contains approximately 25,000 images with 16-keypoint annotations covering a wide range of daily activities. The Human3.6M dataset [39] provides over 3.6 million frames with precise 3D joint annotations captured under controlled multi-camera laboratory conditions. The PoseTrack21 [67] dataset extends multi-person pose estimation to video tracking, while the Yoga-82 [68] dataset serves as a domain-specific benchmark for fine-grained pose classification. Large-scale datasets such as AI Challenger further improved diversity for multi-person pose estimation research [48].

HAR sensor datasets. The Opportunity++ dataset [40, 60] is particularly relevant, as it contains synchronized recordings from both wearable IMU sensors and ambient object-interaction sensors, enabling joint evaluation of multimodal fusion and hierarchical activity recognition involving locomotion primitives and high-level daily activities. The PAMAP2 dataset [41] provides wearable inertial recordings from nine subjects performing 18 physical activities using sensors on the hand, chest, and ankle. The Daily and Sports Activities Dataset (DSADS) [42] contains recordings from eight subjects performing 19 activities using sensors placed on five body segments. Video-based HAR benchmarks such as UCF101 [69] have also been widely used for evaluating activity recognition under unconstrained real-world conditions. Table 2.1 summarizes the key characteristics of these datasets.

Evaluation protocols. For HPE, the standard evaluation metrics include mean Average Precision (mAP) with Object Keypoint Similarity (OKS) on COCO, Percentage of Correct Keypoints (PCK) on MPII, and Mean Per Joint Position Error (MPJPE) in millimetres on Human3.6M [29, 37–39]. The three most relevant classification metrics for sensor-based HAR are accuracy, precision, and recall. The macro-averaged F1 score is suggested as the most appropriate measure of performance because many imbalanced datasets, such as Opportunity++, contain a majority of samples that belong to the Null

class [5, 40]. Cross-user generalization is typically evaluated with the use of LOSO cross-validation [22]. This method is based on training a model using all available subjects except for one that is used as a "hold out" subject to evaluate performance. The evaluation of performance is averaged across all folds of subjects.

Table 2.1: Benchmark datasets used for HAR and HPE

Dataset	Modality	Subjects	Classes / Keypoints	Annotation Type	Usage
Opportunity++ [40]	Wearable + Ambient	4	Loco + HL	Dual-level	Multimodal HAR
PAMAP2 [41]	Wearable IMU	9	18	Single-level	Sensor HAR
DSADS [42]	Wearable IMU	8	19	Single-level	Sensor HAR
Custom Dataset	Vision (skeleton)	–	4	Single-level	Skeleton HAR
COCO [37]	Vision (image)	–	17 keypoints	Keypoint annotations	Multi-person 2D HPE
MPII [38]	Vision (image)	–	16 keypoints	Keypoint annotations	Single-person 2D HPE
Human3.6M [39]	Vision (3D MoCap)	11	3D joints	3D joint annotations	3D pose estimation
PoseTrack21 [67]	Video sequences	–	17 keypoints	Keypoint + tracking	Temporal pose tracking
Yoga-82 [68]	Vision (image)	–	82 classes	Class labels	Fine-grained pose classification

2.7 Research Gaps

The review of the existing literature reveals several important gaps that collectively define the research space addressed in this study.

Gap 1: Absence of a unified comparative benchmark for DL-based HPE. The models proposed and tested are not integrated into a single framework for a fair comparison. The numbers across studies are rarely comparable: training data, augmentation strategies, and hardware all vary, sometimes significantly. To address this, this research aims at developing a structured cross architectural comparison, based on layers, datasets, and metric class, as described in [29, 30].

Gap 2: Lack of standardized pose backend comparison for skeleton-based HAR. Most existing HAR studies that use skeleton-based inputs rely on a single pose estimation backend and do not analyze how backend selection affects downstream recognition accuracy, latency, and efficiency [32]. This makes it hard to choose a backend for a given deployment scenario. This gap is addressed by evaluating multiple pose estimation backends under consistent architectural and experimental conditions.

Gap 3: Limited cross-user generalization in multimodal sensor-based HAR. Sensor-based HAR models often perform well when tested on the same subjects they were trained on, but accuracy drops substantially when applied to new users, a result of variability in how different people execute activities, place sensors, and move [5, 22].

Few studies simultaneously address modality fusion, hierarchical activity structure, and subject-independent evaluation within a single framework [34, 35]. This gap is addressed through a multimodal framework evaluated under LOSO conditions.

Gap 4: Insufficient integration of temporal modeling depth, multimodal fusion, and deployment constraints in resource-constrained environments. CNN and LSTM-based approaches capture local and medium-range temporal dependencies effectively but are limited in modeling complex activity transitions that span extended time horizons [52, 70]. Transformer-based temporal modeling improves long-range dependency modeling but introduces computational overhead that limits deployment in resource-constrained environments [35, 66]. No existing framework brings attention-based multimodal fusion, transformer-based temporal encoding, hierarchical activity modeling, and deployment efficiency together in one place. This gap is addressed through an integrated architecture combining these components. Table 2.2 maps existing work against these gaps.

Table 2.2: Positioning of existing works with respect to identified research gaps

Work	Gap 1	Gap 2	Gap 3	Gap 4
Chung et al. [30]	Partial	×	×	×
Debalaxmi et al. [32]	×	Partial	×	×
Belal et al. [71]	×	×	Partial	Partial
Xu et al. [54]	×	×	Partial	×
Xaviar et al. [34]	×	×	Partial	Partial
Khan et al. [35]	×	×	Partial	Partial

Partial = Partially addressed, ×= Not addressed

2.8 Summary

This chapter covered the literature on HPE, vision-based and sensor-based HAR, multimodal fusion, benchmark datasets, and evaluation protocols. The broad arc is clear: handcrafted features gave way to DL, and CNNs, LSTMs, GCNs, and transformers each pushed recognition performance further. But the literature also has consistent blind spots. Pose estimation backends are rarely compared head-to-head for skeleton-based HAR. Cross-user generalization in multimodal systems remains poor. And temporal modeling, multimodal fusion, and deployment efficiency are almost never tackled together in a single framework. The identified research gaps establish the motivation for the proposed methodologies presented in the subsequent chapters.

CHAPTER 3

COMPARATIVE ANALYSIS OF POSE ESTIMATION MODELS

This chapter introduces a systematic comparison of DL approaches for HPE. The emphasis is on a representative study of pose estimation models for various architectural families and testing their accuracy, computational cost and feasibility in real-time. It examines benchmark datasets and evaluation metrics side by side, both in terms of what they measure, but more importantly, how the selection of a benchmark impacts on the reporting of good performance. Then it discusses accuracy versus cost and correlates attributes of models to environments: Health Care, Fitness, and Surveillance. The results are directly used in the skeleton-based HAR framework described in Chapter 4.

3.1 Overview of DL-based Pose Estimation Approaches

Over the years, DL has generated many different architectures of HPE, that make various compromises between accuracy, speed and deployment constraints [29, 43]. To frame the comparative analysis, this section briefly summarizes the three main categories, rather than repeating the methodological history, as in Chapter 2. CNN-based models extract hierarchical spatial features and achieve good results on the typical benchmark [24, 28, 44]. Lightweight models compromise a bit on accuracy, but are suitable for real-time inference and low compute resources for edge devices [13, 20, 72]. Transformer based-models take advantage of the self-attention mechanism to model long-range spatial relationships, thus being effective in more challenging cases such as occlusion or multi-person cases [25, 26, 49, 73]. The three categories vary significantly in both accuracy and computational expense, and also in the scalability; these points are explored in the following sections.

3.2 Taxonomy of Pose Estimation Architectures

Pose estimation models differ in ways that matter for deployment: how accurate they are, how much compute they need, and how well they hold up outside controlled conditions [29, 43]. These dimensions structure the comparative analysis in this chapter. Figure 3.1 shows the taxonomy used in this study.

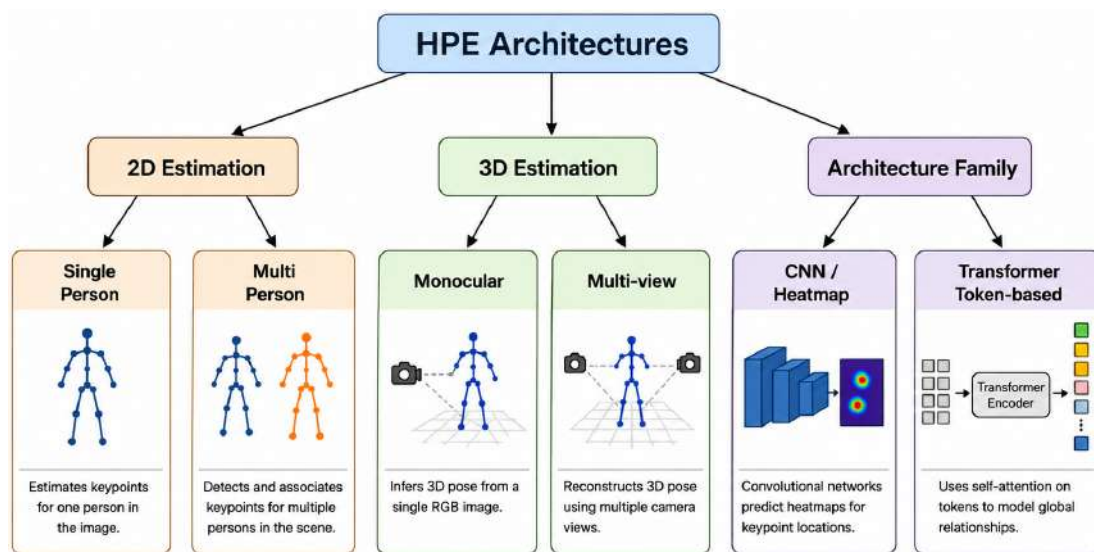


Figure 3.1: Taxonomy of HPE architectures

2D vs. 3D estimation. Estimation of the joint locations in the image plane is simpler and faster using 2D methods, which are adequate in most activity recognition and fitness monitoring applications. [10, 74]. The spatial joint coordinates can be recovered using 3D methods, which can capture the depth and limb orientation information not captured by 2D methods; however, it requires more complex modeling and is prone to failure in non-controlled environments [25, 26, 39].

Single-person vs. multi-person estimation. Single-person methods ignore other subjects, and only try to refine the localization of the keypoints of a single subject. Multi-person methods detect and relate keypoints simultaneously in multiple persons, either using a top-down or bottom up approach. One of the well-known examples of bottom-up open pose is OpenPose [24, 48].

CNN-based vs. transformer-based architectures. CNN-based models are trained hierarchically and are very efficient in practical use in real-time or resource-limited environments. [28, 72]. Self-attention over joints is used in transformer-based models to capture long-range relationships and enables models to handle occlusion and crowded scenes at the cost of higher computational complexity [25, 49, 73].

These dimensions can be combined to provide a unified framework for evaluating accuracy, efficiency, and suitability of pose estimation models for HAR deployments. These dimensions will be combined to establish a consistent framework for evaluating pose estimation models in terms of accuracy, efficiency, and suitability

for HAR deployments. There is no model that has the best performance on all three dimensions. A transformer-based model can work well in occluded/crowded scenes, but may be too expensive to deploy on the edge. A lightweight model may be fine on a small machine, but crash if several users are in frame. These dimensions are then translated to form models that reveal the trade-offs and offer a structured basis for the comparative analysis [12, 15, 33].

3.3 Benchmark Datasets and Evaluation Metrics: Comparative Perspective

Behavior of the models varies across scenarios and is influenced by scale, annotation type and environmental diversity. Even when a model achieves similar results on a benchmark task, it may differ in its ability to handle occlusion and background clutter when tested on large-scale data that has not been used to train the model. Benchmark datasets therefore should be interpreted within the context in which they are used for training and evaluation, rather than the numbers they generate, [26, 29].

There are a number of benchmark sets available for use in HPE, and they vary in complexity and application domain. COCO and MPII focus on 2D pose estimation in real-world settings, such as occlusion and interactions between multiple people. Human3.6M provides highly accurate 3D annotations, but captured in a controlled lab [37–39]. PoseTrack21 also evaluates temporal sequences, where the consistency of the keypoints across video frames is required. Yoga-82 is a domain-specific, fine-grained classification of postures, designed for yoga and related activities [67, 68, 74]. Multi-person pose estimation research was further extended to large-scale datasets like AI Challenger, which added diversity of scenes and scale of annotation [48]. These differences are manifest in the performance of the models. The models trained with COCO are more robust in cluttered multiple-person settings. An evaluation based on MPII favors for precise localization of a single person. When tested on Human3.6M, models obtain low level of spatial errors which is attributed to the attached environment, but this level of accuracy may not be applicable to real-world scenarios [26]. They become better in their domain of knowledge (Yoga-82) but less generalist in other domains.

Evaluation metrics. The nature of the data set is closely related to evaluation in HPE. The standard metrics for 2D pose estimation are PCK and mAP, whereas for 3D tasks, it is MPJPE [29, 43]. Temporal datasets also include the F1 score as an additional metric to evaluate the consistency of the tracking over the various frames [67]. There is a significant impact on the results of comparative analysis based on the choice of

the dataset and the metric used. Due to different task formulation between COCO and MPII, a model that achieves high mAP on COCO might not be the same model when evaluated on MPII based on PCK. A good Human3.6M performance does not necessarily mean robustness in unconstrained environments. The interactions between these make it important to select an appropriate pose estimation model, based on the characteristics of the datasets, evaluation metrics and application requirements; numbers from different benchmarks do not necessarily mean the same thing.

Table 3.1: Evaluation metrics used in HPE research

Metric	Description	Application
PCK	Percentage of joints within a normalized threshold from ground truth	2D HPE on MPII and COCO
mAP (OKS)	Mean precision using OKS	Multi-person 2D HPE on COCO matching
PCPm	Percentage of correctly predicted limb segments	Full-body articulation analysis
MPJPE	Mean Euclidean error per joint in millimeters	3D HPE on Human3.6M
AUC	Area under the PCK curve across thresholds	General 2D benchmarking
NME	Normalized mean Euclidean error	Face and body pose estimation
F1	Harmonic mean of precision and recall	Pose tracking and temporal evaluation

3.4 Comparative Performance Analysis of Pose Estimation Models

A detailed comparative study of the most popular HPE models is given using benchmark results from the literature [29, 43]. We consider CNN-based and transformer-based model families, measure the results on the COCO validation set using mAP with OKS [37]. As shown in Table 3.2, all the models follow a similar trend of improved accuracy from one generation to another, starting with the early CNN models like OpenPose, and progressing to the transformer-based models like ViTPose [24, 25].

This trend is a sign of the progress made in modern pose estimation architectures towards global representation learning instead of localized feature extraction such as [73] and [49].

Table 3.2: Temporal progression of representative HPE models on the COCO validation set

Model	Year	Backbone	mAP (val)	Params
OpenPose [24]	2017	CNN + PAFs	61.0	25–135M
SimpleBaseline [25]	2018	ResNet-152	73.7	Large
MSPN [29]	2019	ResNet-152 + MSPN	76.1	Large
HRNet-W48 [28]	2020	HRNet-W48	76.3	63.6M
TokenPose-B [29]	2021	ResNet-50 + Token	75.8	Moderate
RTMPose-L [29]	2023	ConvNeXt-Large	78.6	Large
ViTPose-L [25]	2023	ViTAE-L (Transformer)	80.6	>300M

This accuracy increase comes at the real expense of parameters and computing. CNN-based models still provide a good compromise between performance and efficiency. With moderate computational cost, HRNet also achieves competitive accuracy, and RTMPose demonstrates that convolutional architectures can achieve transformer-level accuracy without the overhead of transformers [28, 29]. Transformer-based models achieve the best accuracy by encoding the global spatial relationship between body joints however, they require more memory and compute resources and limit real-time applications [25, 49]. Figure 3.2 shows clear trade-off in predicting the accuracy vs. parameters scale for representative architectures.

In all these results, accuracy and complexity go hand in hand, with outcomes varying according to context of use. The high-capacity transformer models are more expensive to use and tend to outperform the benchmark models, but cannot be effectively used in realtime [25, 49]. Although large-scale benchmarks, like COCO, exclude these lightweight CNN-based approaches, they still beat large-scale models in real-time scenarios due to the benefits of low latency and high efficiency in edge devices, as seen in MoveNet and MediaPipe [13, 20, 72]. This implies that accuracy is not a good criterion for model selection. The needs of an embedded sensor or mobile system that

must always be running are different than those of a video processing system that is off-line. The choice of which system to match those requirements and not just picking whichever system gets the best marks on a benchmark determines whether a system will work in deployment [29].

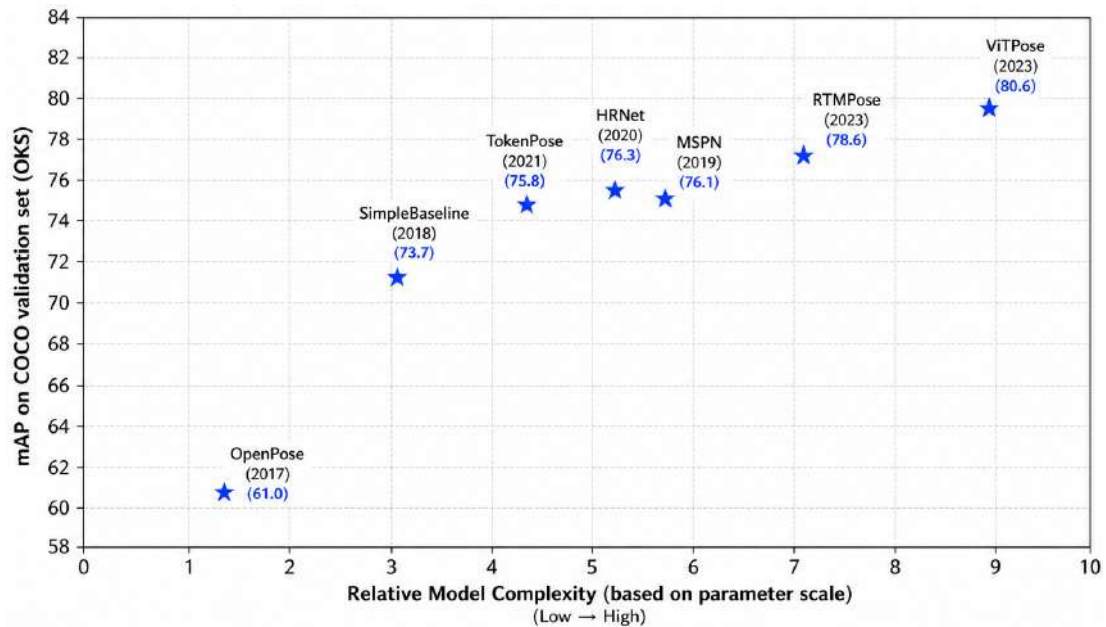


Figure 3.2: Accuracy-complexity relationship of representative HPE models on the COCO validation set (mAP vs. relative model complexity)

3.5 Accuracy and Computational Cost Trade-offs

There is no single HPE architecture that is optimal for all three optimization aspects of accuracy, computational cost and deployment efficiency. All the existing models are in different regions of a trade-off triangle formed by these competing goals, as illustrated in Figure 3.3.

CNN-based models have been developed to achieve a trade-off between accuracy and compute with neither edge, however, they are suitable for applications requiring moderate accuracy and compute capacity, like HRNet [28] and OpenPose [24]. Lightweight models like MoveNet and BlazePose are aimed at the efficiency side of the spectrum: low latency, low memory usage, real-time inference on edge devices. While highly efficient, this also compromises the accuracy of the pose estimation, especially in challenging scenarios such as occlusion or low-light conditions [13, 20, 72]. Transformer-based models work in opposite ways. TokenPose and ViTPose produce the best accuracy by capturing global spatial relationships among different joints, but

they are also complicated and expensive to compute and store, and are hard to deploy in real-time or resource-limited settings [25, 29, 49]. RTMPose sits in between the extremes, and is an optimized convolutional design that achieves transformer level performance but maintains a computational cost closer to CNN level it falls in the middle of the trade-off spectrum [29].

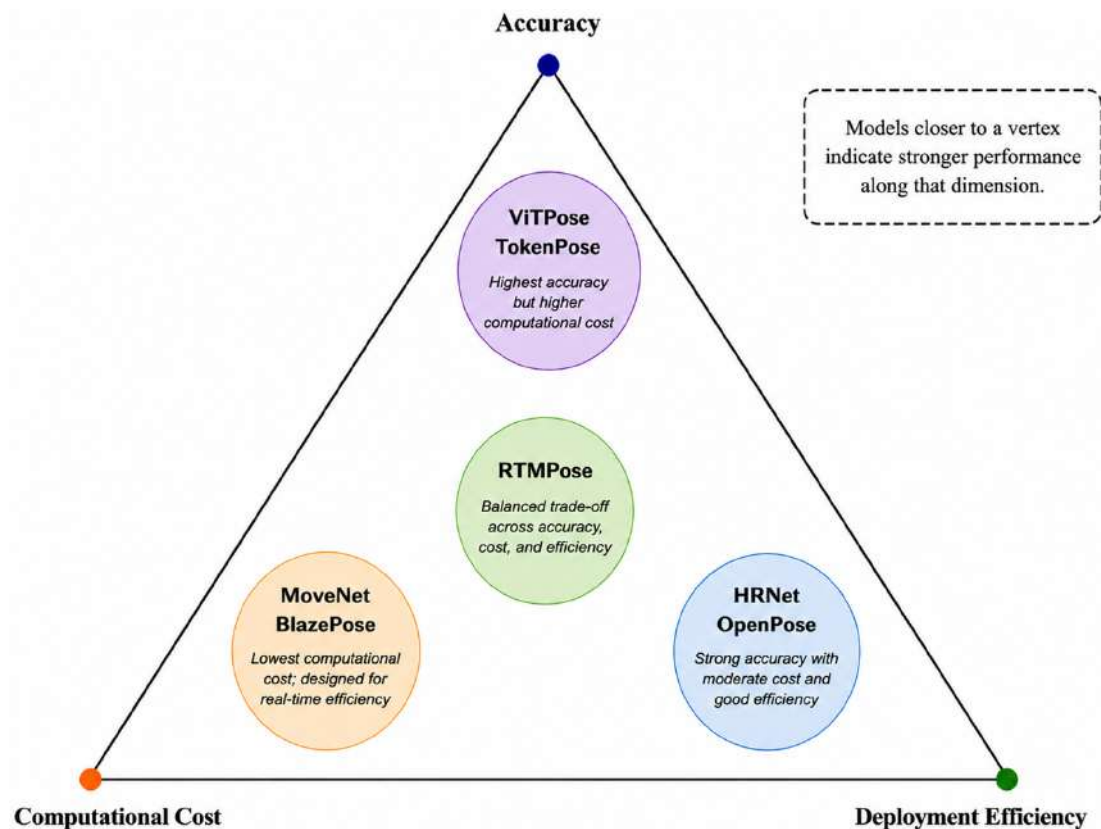


Figure 3.3: Trade-off among accuracy, computational cost, and deployment efficiency in HPE models

Benchmarks are not as relevant to the balance between keypoint accuracy and inference speed for skeleton-based HAR. For real time systems on edge devices, a less accurate pose estimation backend with significantly less latency might be a better fit, depending on the application context they are deployed in and the tasks they must perform. This is one of the reasons for the comparative performance analysis of the different pose backends presented in the following chapters [12, 15, 33]. There is a different way to separate model families for occlusion handling. CNN-based models are limited by their receptive field, meaning they are not certain about prediction when the joint is occluded because they don't see the enough scene to reason about a hidden limb. Transformer-based models build on this with global attention mechanisms, yet

the benefits of robustness are paid for with significant computational expense [25, 49]. It is implied throughout all of the architectural families reviewed here that the accuracy best numbers do not dictate model selection for HAR: there are constraints from the deployment environment. Numbers on a benchmark do not indicate whether a model can run in a memory budget, whether there is a latency requirement, or whether it can be run with partially hidden joints. The constraints are what dictate the operation of a system in the real world.

3.6 Application-based Analysis of Pose Estimation Models

HPE deals with a variety of domains and the specific requirements for each one are sufficiently diverse such that a robust pose estimator does not work well for all of them [29].

Healthcare and rehabilitation. Pose estimation can be used in clinical and rehabilitation applications for non-invasive tracking of patient movements and exercise performance. As for real-time posture tracking and feedback, lightweight models like MoveNet and MediaPipe BlazePose can be implemented on edge devices without special hardware. In such applications, one patient, one camera is almost always the case, and the accuracy of the joints' position and time are much more important than the capability of tracking a crowd [9, 10, 74].

Fitness and sports analytics. Pose estimation is employed for biomechanical analysis, correction of technique and injury prevention. For controlled environments, high accuracy models like HRNet and transformer based models can be feasible. For consumer use, lightweight models are the solution as real-time feedback is the key [25, 28]. Continual tracking of keypoints between frames is crucial, both for motion based metrics and for evaluating the movement pattern, and requires consistent temporal output. The usage scope has also expanded beyond the obvious: for instance, underwater swimmer motion analysis [75] has been one of the applications that has been explored.

Surveillance and crowd monitoring. The process of watching over a particular area or group of people. The practice of monitoring people or groups for surveillance. Different stress applied to the system through surveillance. When multiple people are present, and they are not obstructed, and their viewpoints are not predictable, the default is crowd density, occlusion, unpredictable viewpoints, not the edge case. OpenPose is capable of scaling to these situations with bottom up solutions. High accuracy models often appear in offline or real-time analysis pipelines, where latency is not as critical as

accuracy [24, 48]. The skeleton-based representations enable downstream tasks such as fall detection and behaviour analysis to be more tractable than when working from raw pixels.

HCI and AR. Low latency pose estimation gesture recognition and interactive application breaks down when tracking is laggy with HCI and AR/VR systems. Here lightweight models like MediaPipe BlazePose and PoseNet are more prevalent as the model operates without special hardware and in real-time on mobile and browser-based platforms [13, 72]. Smooth temporal tracking and stable frame rates are not features that one would want to have in these applications, they're essential to the usability of the experience.

Robotics. Pose estimation is a significant aspect of robotics, particularly for human-robot interaction and understanding of robot motion in dynamic settings. Transformer-based models excel at providing spatial reasoning in complex scenarios, however, due to their high computational complexity they cannot be used in lower-end systems without compromising the accuracy of the model. Transformer-based models provide robust spatial reasoning capabilities in complex scenarios, but due to their high computational cost, they are only practical for high-end devices [25, 49]. In most other domains the reliability and real-time constraints are not as relevant as when robots interact with humans, a failed prediction is a safety problem, as well as an accuracy problem.

3.7 Limitations and Future Research Directions

Despite the significant improvement DL has taken HPE, there are still remaining challenges for the performance of these models in the real world [29].

Occlusion and visual variability. Even in unconstrained environments occlusion remains one of the more difficult problems. However, even when the joints are partially hidden, the accuracy and stability of keypoint predictions are compromised, and the severity of occlusion cases are not completely addressed by transformer-based models, although they excel in contextual reasoning [25, 49]. The pressure is compounded by motion blur and low light conditions, as well as sensor noise in real deployments, which is not the case in this experiment and is common in most scenarios [26].

Dataset bias and generalization. Benchmark datasets tend to be focused on controlled environments and not reflective of the diversity of demographics that are found in real use. COCO and Human3.6M trained models tend to perform poorly in out-of-domain environments like healthcare monitoring or surveillance or in outdoor environments,

since they don't match the distribution in which they're trained [37, 39]. Transferring data across different domains is more challenging than it is within a specific domain, while it is easier within this particular domain compared to other ones [26, 68].

Accuracy-efficiency trade-off. The gap between high-accuracy models and models fast enough for real-time deployment has not closed. The highest accuracy is achieved by transformer-based designs, but they consume compute resources that are incompatible with edge use. CNN-based models are lightweight and run in real-time, but are less accurate [13, 20, 25]. Nothing yet exists that sits between them that is comfortable and bridging the gap between them is one of the more persistent practical problems in the field.

Fairness and representation. Performance is not the same across the various demographic groups. Training data imbalance induces biases, not only in accuracy but also in the reliability of systems that humans rely on, as the variations in appearance, body structure and environmental conditions introduce variance [76]. To overcome this, more diverse sets of data and evaluation protocols need to be available, in which demographic variability is an integral part of design, rather than an afterthought.

Future research directions. There are a number of avenues to explore. Datasets must be more extensive, more varied and must be temporally annotated. Model architectures must be more efficient and still improve the accuracy benefits that justify their complexity. Temporal consistency between frames is important for video based pose estimation. One way to increase the diversity of training data is through the use of synthetic data, which can be generated without the challenges of manual annotation, as shown in [25, 26]. Compared to the importance in practice, the evaluation and model design aspects of fairness are underexplored [76].

The identified limitations here show some direction efficient and robust pose estimation frameworks for downstream tasks in real deployments. Inspired by that need, the research of lightweight pose estimation backends for skeleton-based HAR, which is presented in Chapter 4.

3.8 Summary

Architectural family is not the only criteria to assess suitability of a pose estimation model for real-world deployment. It is difficult to compare results across models as the model performances are constrained by the datasets and by the nature of the datasets. New architectures with improved accuracy results in certain test conditions do not necessarily

translate into improved accuracy in all application scenarios. These differences are significant when selecting backends for skeleton-based activity recognition as they are downstream tasks with strict latency and resource requirements for inference. The design decisions and pose estimation options given in the framework introduced in Chapter 4 are natural by-products of this understanding.

CHAPTER 4

SKELETON-BASED HAR USING CNN-LSTM

This chapter introduces a skeleton based HAR system, where multiple pose estimation backends are incorporated into a DL architecture. The comparative analysis of pose estimation models in Chapter 3 is built upon, and the impact of the various pose estimation backends on the performance of activity recognition, while keeping other factors constant, is investigated [30, 32]. A structured pipeline framework is followed: pose estimation, keypoint extraction, feature engineering, temporal modelling and classification. Four models, MediaPipe Pose [13], MoveNet Thunder [20], OpenCV COCO [24] and YOLOv11 Pose [31, 77] were used to extract human skeletal points and converted into structured feature representations covering normalized joint coordinates, inter-joint Euclidean distances, and angular relationships between joints [1]. The frames are grouped into fixed length frame sequences for temporal dynamics. These sequences are then fed into a dual-branch CNN-LSTM fusion network [15, 50] which learns about structural relations from engineered features, while simultaneously considering spatial relations from raw coordinate representations. The fused output is further fed to the recurrent layers to model the temporal information and classify the data. Four daily activities (sitting, standing, walking and running) are evaluated in a standardized train-validation-test split. All backends are evaluated for both classification and computation metrics: accuracy, precision, recall, F1-score, latency, frames per second (fps), model size. This enables the systematic accuracy-by-efficiency comparisons of the various backends that are added [78, 79].

4.1 System Architecture

The primary processing stages of the proposed skeleton-based HAR framework and their interconnections are described in this section. The CNN-LSTM classification pipeline makes use of sequential data from pose estimation and feature generation up to temporal sequence learning and then to activity prediction, all within one.

4.1.1 Overall Pipeline Description

The system architecture consists of a structured pipeline from pose estimation to feature extraction, temporal modeling and activity classification. The outcome of each stage and error propagation from one stage to the other heavily rely on the pose estimation backend to the front end. The overall pipeline and the connectivity of the pipeline's stages are illustrated in Fig. 4.1.

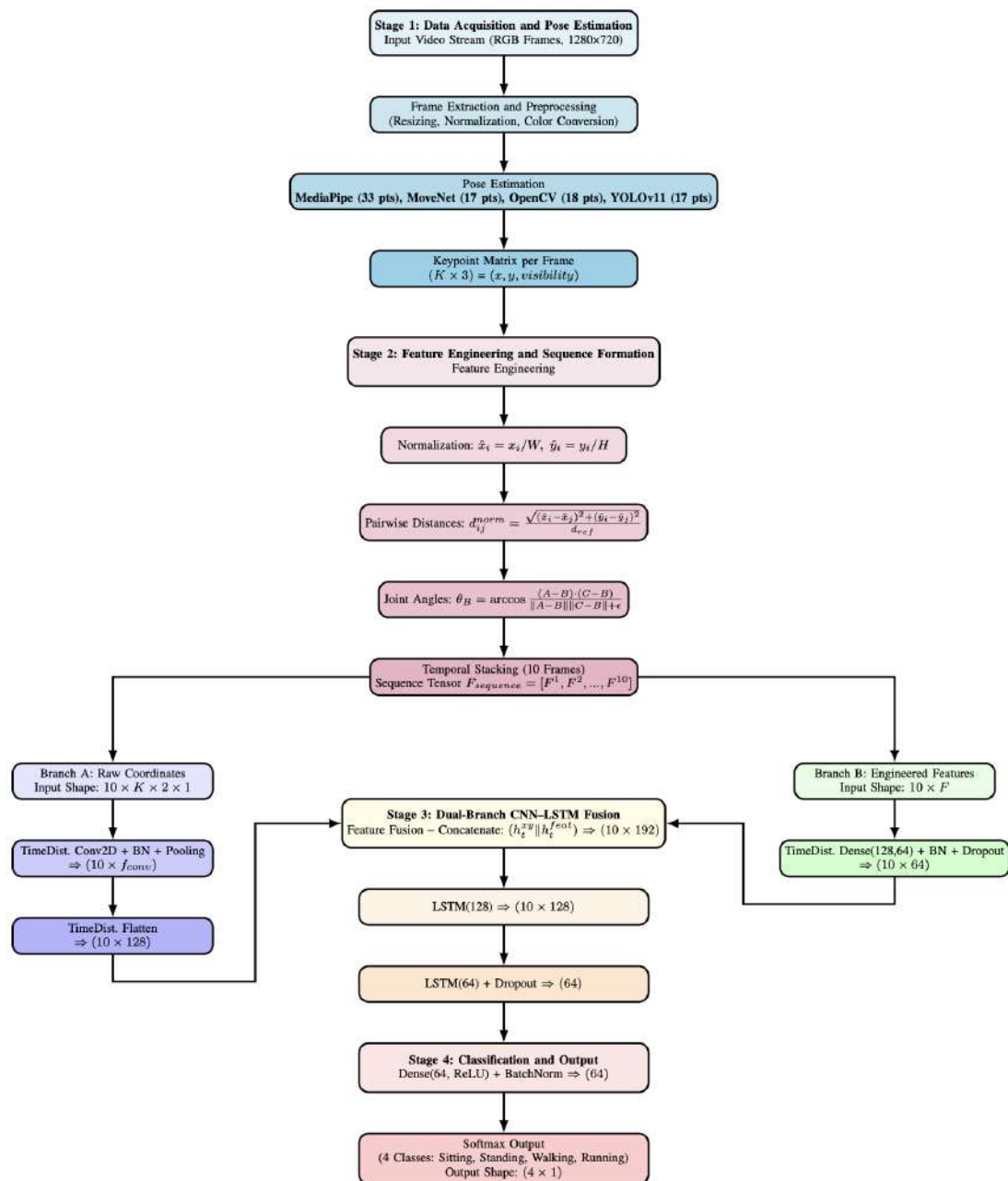


Figure 4.1: Pipeline of the proposed skeleton-based HAR architecture

The first input to the pipeline is a set of input video sequences of human activities. Each frame is extracted and preprocessed based on the input requirement of the pose estimation backends, and then sent to a pose estimate model to get the keypoints of the pose's joints representing the human joint locations. These keypoints are then converted into a set of structured feature representations that include the normalized joint coordinates, inter-joint distances, and angular relationships between adjacent joints [14, 43]. To model temporal dynamics, consecutive frames are purposefully clustered together into fixed-length sequences, so that the evolution of motion can be handled over time, instead of being considered as snapshots. The sequences are passed through a dual-branch CNN-LSTM architecture [15, 50] which is able to learn spatial and temporal dependencies across frames. One is based on raw coordinate representations, the other on engineered features representing structural relationships. The outputs are fused into a single representation, and then run through fully connected layers which classify the activity into one of four pre-defined categories: sitting, standing, walking, or running. All backends share the same downstream architecture with the same experimental conditions, to ensure that the difference in recognition performance and computational cost might be due to the pose estimation stage, not in other parts of the pipeline.

4.1.2 Pose Estimation Backends

For skeletal keypoints extraction, four pose estimation backends were combined into the proposed system, which included MediaPipe Pose [13], MoveNet Thunder [20], OpenCV COCO [24], and YOLOv11 Pose [31]. These models are various compromises between the density of keypoints, computational efficiency, and robustness. The following Table 4.1 summarizes the main features of the backends chosen.

Table 4.1: Pose estimation backends used for keypoint extraction

Model	Keypoints	Output	Framework	Key Strength
MediaPipe Pose	33	(x, y, v)	MediaPipe	High spatial precision
MoveNet Thunder	17	(y, x, c)	TensorFlow Hub	Real-time efficiency
OpenCV COCO	18	(x, y, p)	OpenCV DNN	Stable baseline
YOLOv11 Pose	17	(x, y)	Ultralytics	Robust under occlusion

MediaPipe Pose [13] provides 33 body landmarks, their normalized coordinates, and their visibility score, providing a detailed spatial description of the human pose. It requires BGR to RGB conversion before inference and handles scaling internally. **MoveNet Thunder** [20] outputs 17 keypoints in COCO format and is built for low-latency inference, with input frames resized to 256×256 using aspect ratio-preserving padding. **OpenCV COCO** [24] takes a convolutional pose estimation approach using PAFs, producing 18 keypoints with associated confidence scores. **YOLOv11 Pose** [31] handles joint detection and keypoint estimation in a single stage and holds up reasonably well under motion and occlusion. Each backend has a different preprocessing configuration, as summarized in Table 4.2.

Table 4.2: Preprocessing specifications for each pose estimation backend

Model	Input Resolution	Preprocessing Steps
MediaPipe Pose	Auto-scaled	BGR $\rightarrow$ RGB conversion
MoveNet Thunder	256×256	Resize with padding normalization
OpenCV COCO	368×368	Blob creation with normalization
YOLOv11 Pose	$\sim 640 \times 640$	Internal preprocessing

4.2 Dataset Construction

In this section, the organization of the activity dataset for the experimental evaluation is described. Distributions of activity samples in the partition of the datasets and distribution of extracted skeletal keypoint representations to the backends for training and testing.

4.2.1 Data Collection and Preprocessing

To assess the proposed skeleton-based HAR framework under various pose estimation backends, a balanced dataset has been built. A balanced dataset was built for the evaluation of the proposed skeleton-based HAR framework using various pose estimation backends. It is a set of time-synchronized video recordings and extracted image frames from four activity classes, gathered keeping the activity classes evenly distributed across the set of videos so as to enable a fair comparison of performance across backends.

Samples were divided in training, validation and test sets at 70:15:15 ratio. OpenCV [80] was used to extract and resize frames, and TensorFlow [81] was used to

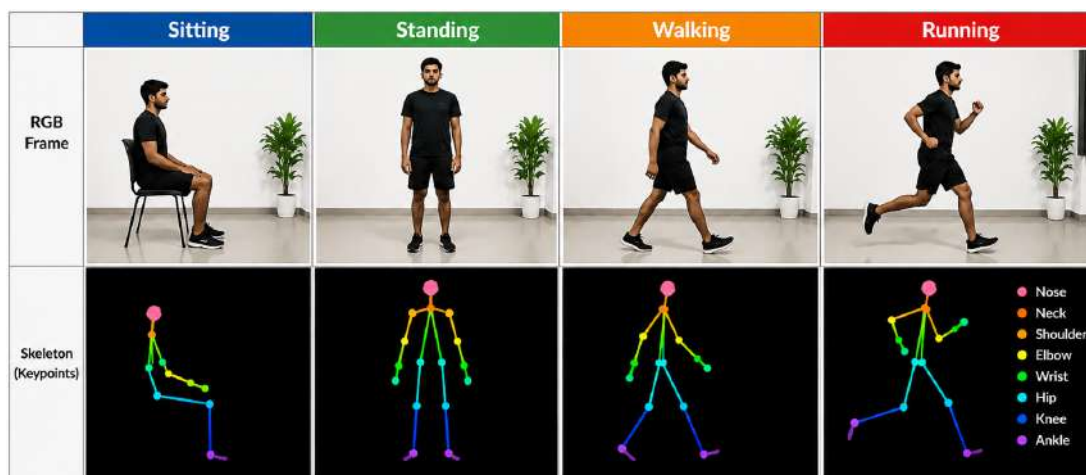


Figure 4.2: Representative RGB frames and corresponding skeletal keypoint representations for the four activity classes

process them for the backend. To extract motion dynamics for activity recognition, each video sequence was divided into fixed size temporal windows of 10 consecutive frames. The 10 frames are enough to give sufficient context to distinguish periodic activities from one another without being too expensive for real-time inference. Some sample images from the dataset are shown along with their respective skeletal keypoint representations in each of the four activity classes in Figure 4.2. The overall characteristics of the dataset are summarized in Table 4.3.

Table 4.3: Summary of the constructed HAR dataset

Property	Value
Activity classes	Sitting, Standing, Walking, Running
Number of classes	4
Data sources	Video recordings and extracted frames
Class distribution	Balanced across all classes
Train / Val / Test split	70% / 15% / 15%
Preprocessing tools	OpenCV and TensorFlow
Temporal window size	10 consecutive frames

A model that has an imbalanced class distribution tends to overfit onto the majority class and thus overestimates accuracy rates in HAR, which is why class balance is important. In addition to class balance, the dataset is organized in a coherent way across

all video samples, extracted image frames and temporally segmented activity sequences during training. This means that each category of activity has a similar spatial and temporal signal throughout the learning and evaluation period, no single class dominates the signal the model sees. Table 4.4 shows the distribution of videos, image frames, and sequence samples between the training, validation and test sets.

Table 4.4: Balanced data distribution across activity classes

Class	Type	Train	Val	Test
Sitting	Videos	76	16	17
	Images	2307	494	496
	Sequences	99	21	22
Running	Videos	76	16	17
	Images	2307	494	496
	Sequences	99	21	22
Standing	Videos	76	16	17
	Images	2307	494	496
	Sequences	99	21	22
Walking	Videos	76	16	17
	Images	2307	494	496
	Sequences	99	21	22

Once preprocessed, both pose estimation backends extracted the skeletal keypoints independently and the extracted keypoints were stored as numerical arrays for feature engineering and temporal modeling. There are different numbers of keypoint samples extracted per backend due to the variation in the confidence of pose detection, the stability of landmark tracking, and the behavior of estimating the keypoints frame-wise. Denser representations of landmarks lead to bigger keypoint sets, whereas lightweight models, suitable for real-time inference, may fail to report low confidence landmarks in difficult motions. The keypoint arrays extracted are a set of coordinates for the joints in each of the frames detected, and they form the main input for the skeleton-based HAR pipeline. The differences between the two backends in keypoint generation directly influence the differences in recognition accuracy and computational efficiency in the

evaluation, and thus the spatial representations the CNN-LSTM architecture learns from and feeds into. Table 4.5 shows the statistics of the obtained keypoint sample across the dataset partitions for all the keypoint backends evaluated.

Table 4.5: Keypoint distribution across pose estimation backends

Backend	Class	Train	Val	Test
MediaPipe	Running	7981	1429	1843
	Sitting	25782	5399	5743
	Standing	25502	5452	5720
	Walking	24525	5286	5392
YOLOv11 Pose	Running	30847	5139	7682
	Sitting	26195	5543	5878
	Standing	26129	5583	5895
	Walking	33802	6319	7393
OpenCV COCO	Running	30036	4939	7394
	Sitting	25892	5426	5768
	Standing	25675	5478	5759
	Walking	32506	6228	6915
MoveNet	Running	30036	4939	7394
	Sitting	25892	5426	5768
	Standing	25675	5478	5759
	Walking	32506	6228	6915

4.2.2 Activity Classes

The dataset includes four activities: sitting, standing, walking and running. These were selected since they cover a helpful range of motion complexity from stillness to full locomotion. Sitting and standing are essentially static, with little variation of the body in the pose. They are periodic and change more rapidly over time walking and running are different. The two types in the data set test the framework on different types of movement, rather than only one corner of the problem.

4.3 Keypoint Feature Engineering

Structured feature representation was used as the input of the CNN-LSTM fusion architecture, and was generated by transforming raw skeletal keypoints extracted by each pose estimation backend. The retrieved keypoints were encoded in a compact representation that encodes the spatial posture and articulation of the limbs. Three sets of features were calculated following the methodology described, which were normalized joint coordinates, inter-joint distances and joint angles. The functionality and feature preprocessing were numerically calculated with the NumPy array operations [82]. In Figure 4.3 the spatial features extracted from the skeletal keypoints are shown as examples of normalized coordinates, inter-joint distances and joint angles.

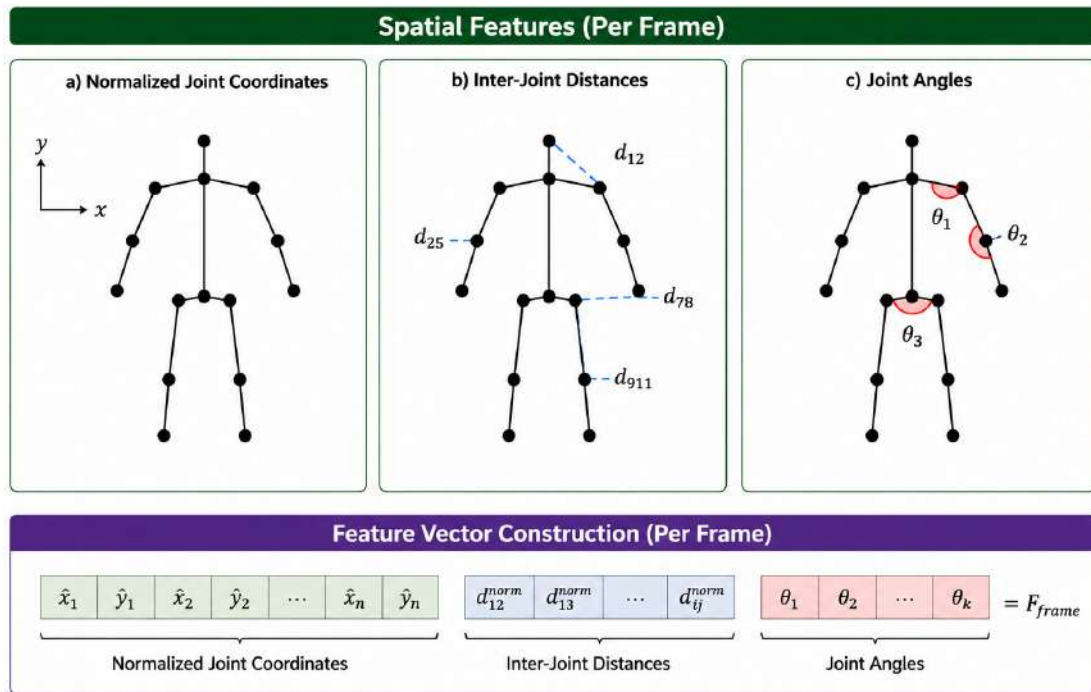


Figure 4.3: Illustration of spatial features derived from skeletal keypoints, including normalized joint coordinates, inter-joint distances, and joint angles

4.3.1 Normalized Joint Coordinates

Normalized coordinates were obtained by dividing each joint position by the frame width W and height H to make pose representation invariant to image scale. This transformation is defined in Eq. 4.1:

$$\hat{x}_i = \frac{x_i}{W}, \quad \hat{y}_i = \frac{y_i}{H} \quad (4.1)$$

where x_i, y_i represent the original joint coordinates and W, H denote the frame width and height.

4.3.2 Inter-Joint Distance Features

Pairwise Euclidean distances between keypoints were computed to capture relative spatial relationships between body joints. These distances were normalized using a reference distance d_{ref} defined between the shoulder and hip joints. The normalized distance is defined in Eq. 4.2:

$$d_{ij}^{\text{norm}} = \frac{\sqrt{(\hat{x}_i - \hat{x}_j)^2 + (\hat{y}_i - \hat{y}_j)^2}}{d_{\text{ref}}} \quad (4.2)$$

where d_{ref} represents the normalization reference distance between the shoulder and hip joints.

4.3.3 Joint Angle Representation

Angular relationships between adjacent joints were computed to represent limb articulation and rotational motion. The angle at joint B as calculated using Eq. 4.3:

$$\theta_B = \arccos\left(\frac{(A - B) \cdot (C - B)}{\|A - B\| \cdot \|C - B\| + \varepsilon}\right) \quad (4.3)$$

where A, B, C are the positions of adjacent joints used to calculate the angle at joint B and ε is a small constant to prevent numerical instability.

4.3.4 Temporal Feature Construction

The final feature vector at the frame level was created by concatenating the normalized coordinates, inter-joint distances and joint angles as shown in Eq. 4.4:

$$F_{\text{frame}} = [\hat{x}_1, \hat{y}_1, \dots, \hat{x}_n, \hat{y}_n, d_{ij}^{\text{norm}}, \theta_{ABC}] \quad (4.4)$$

In order to achieve temporal continuity, the feature vectors of consecutive frames were concatenated into fixed length sequences. In particular, 10 successive frames were composited together to create a spatio-temporal tensor, as described in Eq. 4.5:

$$F_{\text{sequence}} = [F_{\text{frame}}^1, F_{\text{frame}}^2, \dots, F_{\text{frame}}^{10}] \quad (4.5)$$

where ten consecutive frames are stacked to capture temporal dependencies across motion sequences. Table 4.6 summarizes the resulting input dimensions for each pose estimation backend.

Table 4.6: Input dimensions for the CNN-LSTM fusion architecture

Model	XY Input Shape	Feature Input Shape
MediaPipe Pose	$10 \times 33 \times 2 \times 1$	10×78
MoveNet Thunder	$10 \times 17 \times 2 \times 1$	10×46
OpenCV COCO	$10 \times 18 \times 2 \times 1$	10×48
YOLOv11 Pose	$10 \times 17 \times 2 \times 1$	10×46

4.4 Dual-Branch CNN-LSTM Fusion Architecture

The classification model is a dual-branch CNN-LSTM fusion architecture designed to jointly capture spatial structure and temporal dynamics of human activities from skeletal keypoints. The architecture is based on a hybrid DL paradigm for sequence modeling that is based on the convolutional feature extraction and recurrent temporal modeling of the sequence [50]. This design enables both pose configuration and motion evolution over time to be learned effectively. The overall architecture of the classification model is shown in Figure 4.4.

4.4.1 Spatial Feature Learning using CNN

A convolutional branch is used to learn the spatial structure of human pose, where raw joint coordinate sequences are fed into this branch. The coordinates of the structured keypoints for each frame are represented, and a TimeDistributed convolutional pipeline is used to extract the spatial features independently at every time step. To capture the spatial relationships between the joints, a sequence of TimeDistributed convolutional layers with Batch normalization and Max pooling is used. The output is a sequence of feature maps which are flattened to get a temporal sequence of feature vectors h_t^{xy} that describe the features.

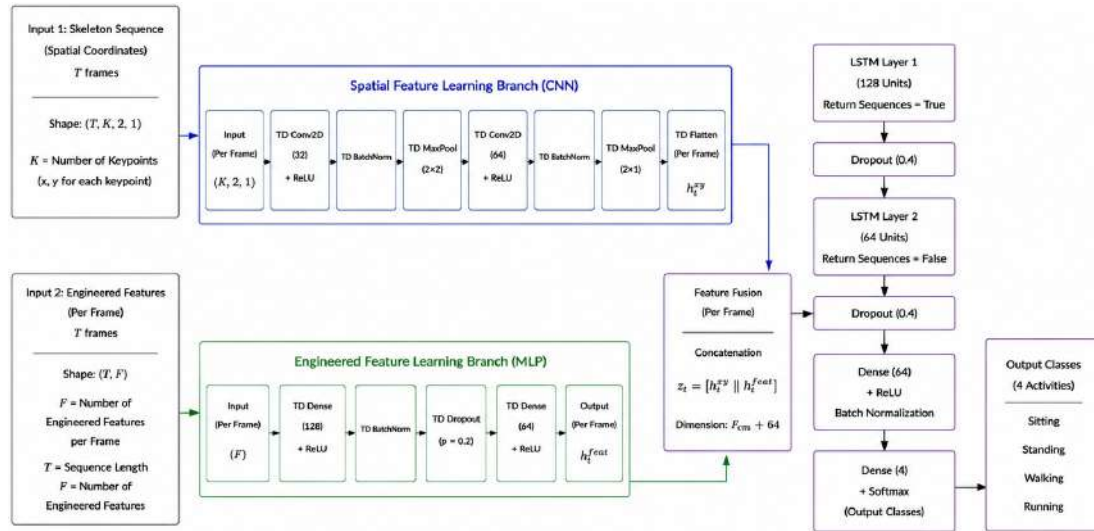


Figure 4.4: Dual-branch CNN-LSTM fusion architecture for skeleton-based HAR

4.4.2 Engineered Feature Learning Branch

The second branch handles engineered features based on the skeletons keypoints, which include normalized coordinates, intra-joint distances and angular relations. These features give an abbreviated encoding of the geometric and kinematic aspects of human motion. At each time step, a TimeDistributed multilayer perceptron with dense layers, batch normalization and dropout layers are used. This branch's output is the feature vector h_t^{feat} which represents the relationship between joints.

4.4.3 Temporal Modeling using LSTM

An LSTM layer is used in a stack to capture the temporal correlations between the next two frames. After fusion, the concatenated feature sequence is fed into 2 LSTM layers to learn the dynamics and sequential information of human activities. The first LSTM layer has 128 units, which is used to process the whole temporal sequence and keep temporal continuity. The final hidden state is passed to a second LSTM layer containing 64 units that produces a compact representation of the overall activity pattern.

4.4.4 Feature Fusion and Classification

To combine both spatial and kinematic information, feature fusion is done at every time step by adding together the two branches output. The fused representation is defined in Eq. 4.6. Concatenated sequence Z_t passed through the stacked LSTM layers. The final hidden state h_T is then passed to a fully connected layer and a Softmax output layer as shown in Eq. 4.7 for activity prediction. There are 4 activity classes that the model is

able to classify: sitting, standing, walking and running.

$$Z_t = \begin{bmatrix} h_t^{xy} || h_t^{feat} \end{bmatrix}, \quad h_t = \text{LSTM}(Z_t, h_{t-1}) \quad (4.6)$$

$$\hat{y} = \text{Softmax}(W \cdot h_T + b) \quad (4.7)$$

4.5 Experimental Setup and Evaluation Metrics

The experimental evaluation was structured to test the performance of each pose estimation backend in two complementary aspects: the classification performance and computational efficiency. HAR systems should have to be capable of high recognition accuracy and also be able to process at real-time speed and hence the evaluation was designed to satisfy both of these demands.

4.5.1 Training Configuration

To facilitate a fair comparison, all pose estimation backends were tested with the same CNN-LSTM fusion architecture, data splitting, and preprocessing. Dataset split into training, validation, and test set, and the hyperparameters were the same in all experiments. A batch size of 64, Adam optimizer with $\eta = 0.001$ and categorical cross-entropy loss were used for optimizing the network. An early-stopping technique by using the validation loss was used to prevent overfitting and guarantee stable convergence. To account for stochastic training, each experiment was replicated three times, and the results reported here are the average of the three runs. The variance of these runs was small; less than 5.0% (plus or minus 0.5%-1.0%), which suggested that the models behaved in a stable and consistent manner. The tests were performed on an AMD Ryzen 7 5800H system with 8 GB of RAM and NVIDIA GeForce RTX 3050 graphics card.

4.5.2 Evaluation Metrics

Each pose estimation backend was tested by both classification and efficiency measures.

Accuracy. Accuracy is the percentage of the samples of the test set that are correctly classified, according to Eq. 4.8:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.8)$$

Precision, Recall, and F1-score. The precision, recall and F1-score offer a detailed measure of the performance of the classification. Precision refers to the accuracy of correctly predicted positive instances, recall refers to the accuracy of the instances that are relevant to the predicted class and F1-score is the harmonic mean of precision and recall. They are defined here as follows in Eqs. 4.9–4.11:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.10)$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.11)$$

Latency. Latency is the average inference time (ms) per frame, or the time needed to infer a single frame.

fps. The fps value is calculated using the equation below, which is the inverse of the latency, as indicated in Eq. 4.12:

$$\text{fps} = \frac{1000}{\text{Latency}(\text{ms})} \quad (4.12)$$

Model Size. The overall memory consumption of the trained CNN-LSTM model (in Megabytes) and the number of model parameters.

All efficiency metrics were taken on CPU to simulate the environment of deployment in resource-constrained devices like edge devices and IoT.

4.6 Results and Comparative Analysis

In this section, the performance of the CNN-LSTM fusion framework, using four pose estimation backends, is presented in the form of a comparison. In skeleton-based HAR, the recognition performance, computation efficiency and the accuracy-

real-time inference behavior are analyzed. By comparing multiple backends, activity recognition and inference performance can be compared to investigate the effect of different keypoint representations on the activity recognition performance and inference speed. The reported results also provide insights in the suitability of each backend to deploy in different application and computation scenarios.

4.6.1 Classification Performance

Table 4.7 shows the classification results of all four pose estimation backends in the dual-branch CNN-LSTM fusion framework. The best performance was seen for MediaPipe Pose with an accuracy of 94%, showing a good balance between precision, recall, and F1 score for each activity class. The 33 landmark representation it uses is responsible for this performance, as it provides information at a finer granularity level and minimizes misinterpretation of similar static poses, like sitting and standing. The improved keypoint structure means greater spatial discrimination for the CNN branch at the time of classification at activity boundaries. YOLOv11 Pose was robust enough to achieve 89% accuracy in dynamic conditions and partial occlusion. Results were unchanging in the higher complexity of motion situations of walking and running. The baseline of OpenCV COCO was 88%. But greater walking/standing confusion was seen, which may be caused by lower keypoint resolution. Landmarks are sparse, and the spatial branch cannot differentiate between the two activity classes, due to fine postural differences. MoveNet Thunder was able to reach an accuracy rate of 85%. The scores obtained for the dynamic activities are acceptable, and the ones obtained for the static poses are comparatively low, due to the low granularity of the keypoints. The results of the classification indicate that the keypoint richness is a meaningful factor for recognition with respect to recognition performance, especially in the case of activity classes with similar postural configurations. The confusion matrices of all four pose estimation backends are presented in Fig. 4.5 and give a class-wise detail of prediction performance.

Table 4.7: Classification performance of the CNN-LSTM fusion architecture across pose estimation backends

Model	Accuracy	Precision	Recall	F1 Score
MediaPipe Pose	0.94	0.94	0.94	0.94
YOLOv11 Pose	0.89	0.88	0.88	0.88
OpenCV COCO	0.88	0.89	0.88	0.88
MoveNet Thunder	0.85	0.86	0.86	0.86

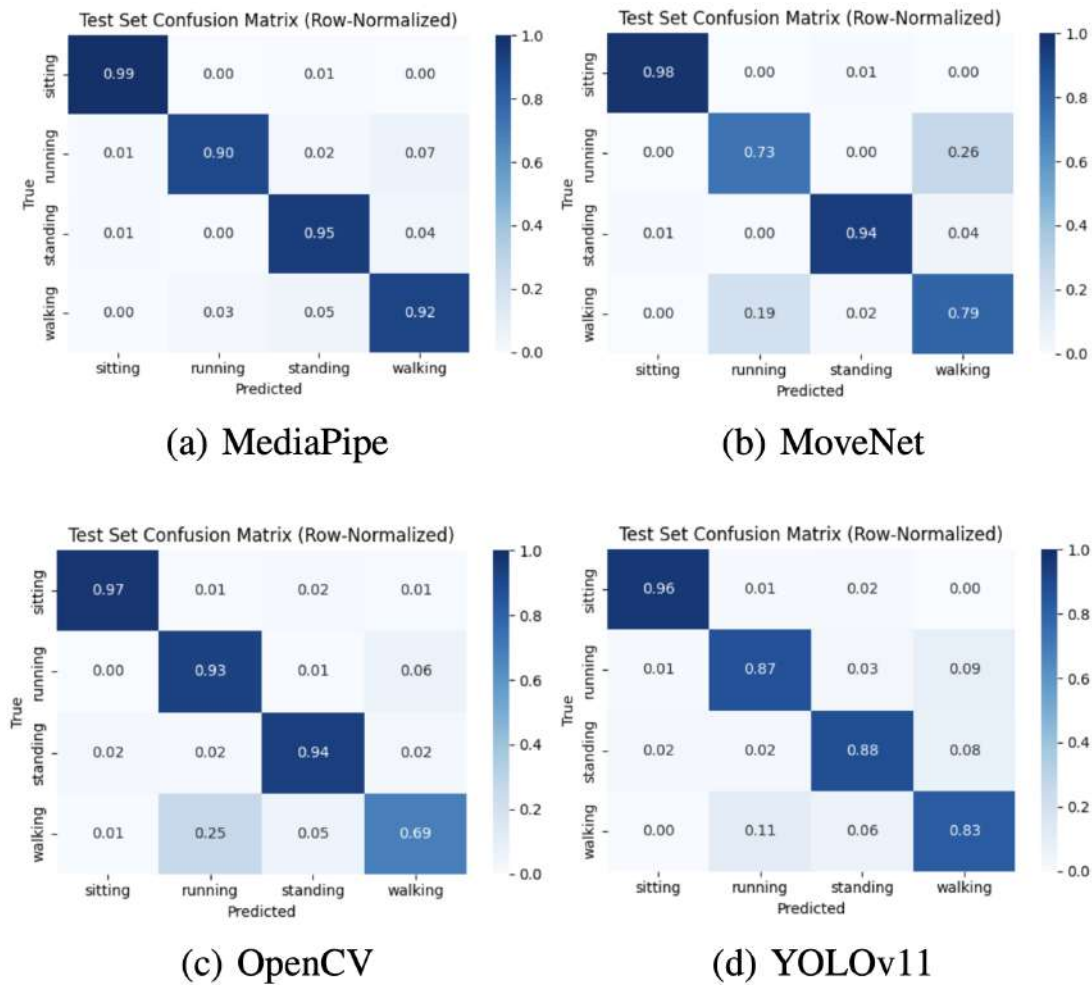


Figure 4.5: Confusion matrices of the four pose estimation backends on the four-activity test dataset

4.6.2 Latency and Throughput Analysis

The most efficient backend for real-time deployment is YOLOv11 Pose with the lowest latency of 125.0 ms and highest throughput of 8.0 fps. YOLOv11's lower dimensionality of keypoint density decreases feature vector size sent to the CNN-LSTM pipeline directly contributing to faster inference without compromising recognition robustness. MoveNet Thunder and OpenCV COCO had similar accuracy, but slightly higher latency and slightly lower throughput. MediaPipe Pose had the highest dimensional representation of keypoints, as well as the highest latency and model size. The 33 landmark structure is useful to keep classification accuracy high, but it introduces more computation in the feature extraction and temporal modeling process which leads to slower inference on the CPU. Table 4.8 shows the computational efficiency results of

the CNN-LSTM fusion model on CPU, for all pose estimation backends.

Table 4.8: Computational efficiency of CNN-LSTM models evaluated on CPU

Model	Latency (ms)	fps	Model Size (MB)	Params (M)
YOLOv11 Pose	125.0	8.0	3.68	0.31
OpenCV COCO	127.4	7.9	3.68	0.31
MoveNet Thunder	128.5	7.8	3.68	0.31
MediaPipe Pose	130.2	7.7	5.23	0.45

4.6.3 Accuracy vs Efficiency Trade-off

Results show a clear trade-off between recognition accuracy and computational efficiency between the considered backends. MediaPipe Pose has the highest accuracy because it represents dense keypoints, yet is the most expensive and has the highest latency. YOLOv11 Pose is the most efficient with both lower latency and higher throughput with competitive accuracy. MoveNet Thunder and OpenCV COCO fall somewhere in between: Moderate accuracy, and relatively efficient inference. The observations highlight that the selection of the pose estimation backend should be based on the application, including the emphasis on real-time performance or accuracy.

4.6.4 Training Convergence

All backends were configured without over fitting and had stable training and validation curves over the epochs. MediaPipe and YOLOv11 converged within 10-15 epochs, with OpenCV COCO taking longer to converge due to the noisier keypoint values. MoveNet Thunder had a fast convergence time but lower performance ceiling. All configurations were trained stably by using Dropout, Batch Normalization, and Early Stopping. The CNN-LSTM fusion architecture exhibited similar learning dynamics across different keypoint representations and feature dimensionalities in the backends used. The training and validation accuracy curves remained largely similar, indicating good generalization for all the tested configurations. In Fig. 4.6 training and validation accuracy are shown for all backends, which shows that the behavior of convergence is stable.

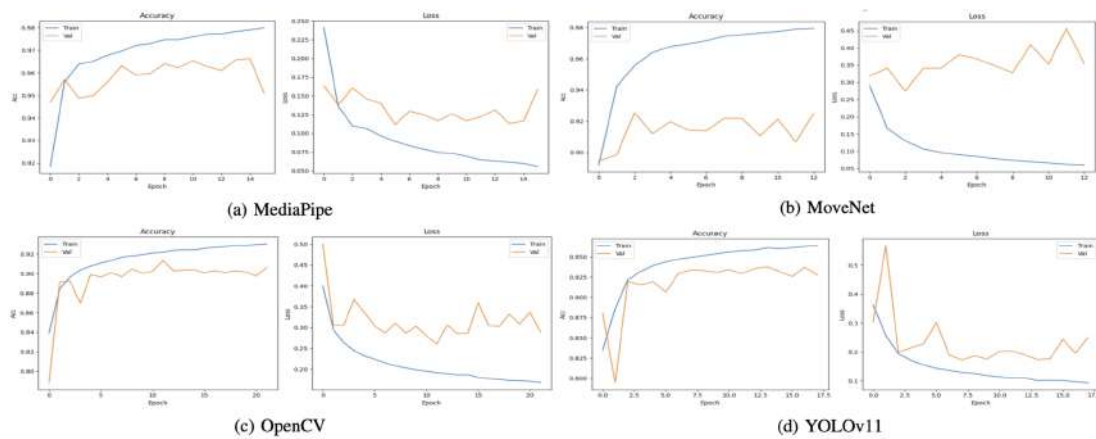


Figure 4.6: Training and validation accuracy trends of the CNN-LSTM fusion model across all pose estimation backends

4.7 Discussion

The experimental result shows that choosing the pose estimation backend greatly affects the recognition accuracy and efficiency in skeleton-based HAR systems. The dense landmark representation of MediaPipe Pose provides the best recognition performance, especially for static activities. YOLOv11 Pose offers better real-time performance, reduced latency and smaller models, which makes it more suitable for deployment on resource-limited devices. The dual-branch CNN-LSTM fusion architecture is found to be robust to the variations of keypoint representation, and thus to be stable across all evaluated backends. CV COCO is an accurate baseline whose performance is balanced. The evaluation is restricted to single person activities in controlled settings, and future work will be focused on multi-person activities, a broad range of activity types and edge deployment optimisation. These observations lead to an interest in the transition to multimodal approaches, described in Chapter 5.

4.8 Summary

This chapter presented a skeleton-based HAR architecture built on multiple pose estimation backends and a dual-branch CNN-LSTM fusion model, covering dataset construction, keypoint extraction, feature engineering, and temporal sequence generation. Experimental evaluation was conducted using MediaPipe Pose, YOLOv11 Pose, OpenCV COCO, and MoveNet Thunder under a unified setup with classification and computational performance metrics. The presented work establishes the basis for the multimodal sensor-based HAR approach discussed in Chapter 5.

CHAPTER 5

MULTIMODAL HIERARCHICAL TEMPORAL LEARNING FOR CROSS-USER HAR

This chapter presents a multimodal learning approach for HAR under cross-user settings using synchronized wearable and ambient sensor streams. The study is motivated by the challenges of inter-subject variability, heterogeneous sensing characteristics, and limitations of existing models in capturing long-range temporal dependencies in multimodal HAR systems [19, 23, 65]. To address these issues, a unified framework is developed that integrates modality-specific convolutional encoders for temporal feature extraction, a learnable scalar attention-based mechanism for modality fusion, and transformer-based temporal modeling to capture dependencies within sensor sequences. The framework further incorporates a multi-task learning formulation to jointly model locomotion-level and high-level activities from a shared representation. The evaluation is conducted under a LOSO protocol on the Opportunity++, PAMAP2, and DSADS datasets to assess generalization across unseen users [27, 60]. In addition, a lightweight architectural variant based on depthwise separable convolution is designed to reduce computational complexity while preserving the multi-task prediction structure for resource-constrained deployment scenarios [21, 36]. The methodology and evaluation presented in this chapter are derived from the conducted research work.

5.1 Problem Formulation and Challenges

Robust HAR under cross-user settings remains challenging due to variability in human behavior and the heterogeneous nature of sensor modalities. These challenges are only partially addressed by existing approaches [5, 23]. In this study, HAR is formulated as a window-level multi-task classification problem over synchronized wearable and ambient sensor streams while addressing key limitations observed in prior work.

5.1.1 Inter-Subject Variability

Inter-subject variability arises due to differences in how individuals perform the same activity. Variations in motion patterns such as walking cadence, stride length, posture,

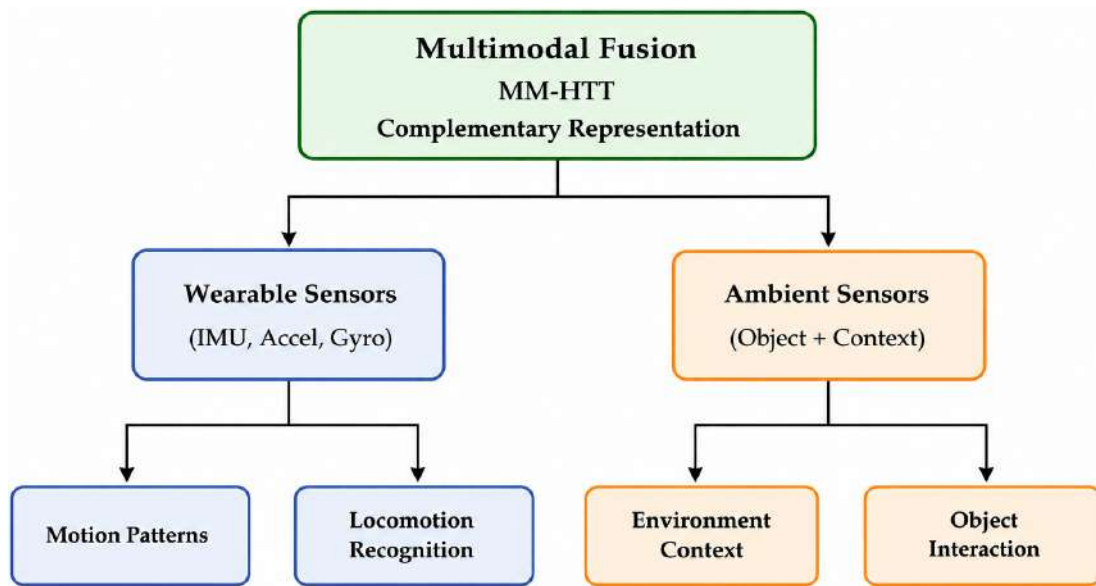


Figure 5.1: Complementary characteristics of wearable and ambient sensor modalities motivating multimodal fusion

and arm movement lead to significant differences in sensor signals across subjects [4,22]. Consequently, models trained on data from specific individuals tend to learn subject-dependent patterns rather than generalized activity representations. This results in reduced performance under cross-user evaluation settings, particularly for activities with high variability in execution.

5.1.2 Sensor Heterogeneity

Another major challenge is the heterogeneous nature of multimodal sensor data. Wearable sensors capture detailed body kinematics through continuous inertial measurements and are effective for recognizing locomotion activities characterized by periodic motion patterns [7, 16]. However, they lack contextual awareness and are insufficient for distinguishing activities that involve similar motion patterns but different environmental interactions. Ambient sensors, on the other hand, provide contextual information through object interaction events and environmental signals [11, 40]. These sensors are effective for recognizing high-level activities that depend on contextual cues but are limited in capturing fine-grained body motion. As a result, neither modality alone is sufficient for comprehensive activity recognition. As illustrated in Fig. 5.1, the complementary strengths of wearable and ambient sensing motivate the integration of multimodal information within a unified learning framework.

The proposed MM-HTT framework formulates multimodal HAR as a window-

level multi-task classification problem over synchronized wearable and ambient sensor streams. Let $\mathbf{X}$ denote a multivariate sensor window of length T , where each time step contains a synchronized feature vector composed of wearable and ambient sensor measurements, as defined in Equation 5.1. The framework learns a mapping function $f_{\theta}(\cdot)$ that jointly predicts locomotion-level and high-level activities from the input sequence, as expressed in Equation 5.2. The final activity labels are obtained from the maximum response of the corresponding task-specific output logits according to Equation 5.3. To jointly optimize both recognition tasks, the framework is trained using the weighted multi-task objective defined in Equation 5.4, where separate loss terms are assigned to locomotion and high-level activity recognition. Each task loss is computed using categorical cross-entropy with label smoothing, and samples corresponding to the Null class are excluded from both training and evaluation to reduce bias caused by class imbalance.

$$\mathbf{X} = \{x_1, x_2, \dots, x_T\} \quad (5.1)$$

$$(o_{\text{loc}}, o_{\text{hl}}) = f_{\theta}(\mathbf{X}) \quad (5.2)$$

$$\hat{y}_{\text{loc}} = \arg \max o_{\text{loc}}, \quad \hat{y}_{\text{hl}} = \arg \max o_{\text{hl}} \quad (5.3)$$

$$\mathcal{L} = w_{\text{loc}} \mathcal{L}_{\text{loc}} + w_{\text{hl}} \mathcal{L}_{\text{hl}} \quad (5.4)$$

5.2 Proposed Multimodal Learning Framework

The proposed MM-HTT framework integrates modality-specific convolutional encoding, learnable multimodal fusion, and transformer-based temporal modeling for cross-user HAR. The architecture is designed to preserve temporal resolution while enabling effective multimodal representation learning for cross-user activity recognition. As illustrated in Fig. 5.2, the framework consists of modality-specific convolutional encoders, a learnable attention-based fusion module, and a transformer encoder for temporal modeling.

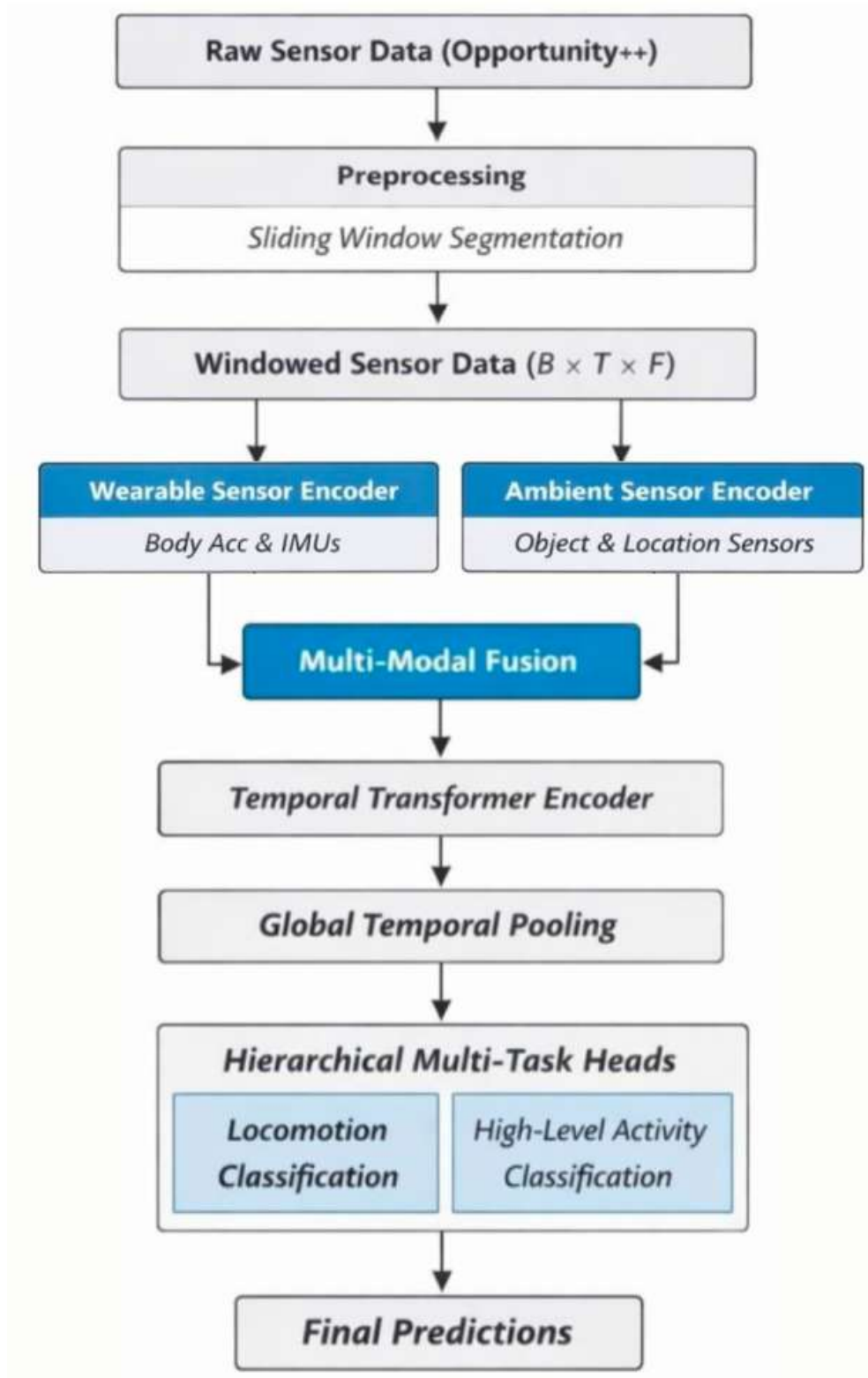


Figure 5.2: Architecture of the proposed multimodal hierarchical temporal learning framework

5.2.1 Multimodal Inputs: Wearable and Ambient Sensors

The framework operates on synchronized multimodal sensor streams comprising wearable and ambient modalities. The wearable stream captures body motion through inertial measurements, while the ambient stream provides contextual information through environmental and object interaction signals. Let $\mathbf{X}_w \in \mathbb{R}^{T \times F_w}$ and $\mathbf{X}_a \in \mathbb{R}^{T \times F_a}$ denote the wearable and ambient input sequences, respectively, where T represents the temporal window length and F_w, F_a correspond to modality-specific feature dimensions. These synchronized inputs are processed independently through modality-specific encoders before fusion.

5.2.2 Modality-Specific Convolutional Encoders

The first stage of the proposed framework consists of modality-specific one-dimensional convolutional encoders, where separate encoders are used for the wearable and ambient sensor streams to extract temporal feature representations independently prior to fusion. Each encoder comprises three sequential convolutional blocks. Each block consists of a Conv1D layer followed by Batch Normalization, ReLU activation, and Dropout regularization with a rate of 0.2. No temporal downsampling is applied within the convolutional blocks, ensuring that the temporal resolution of the input sequence is preserved. This design allows the subsequent transformer encoder to operate on complete temporal feature sequences and model dependencies across all time steps.

For each modality $m \in \{w, a\}$ corresponding to wearable and ambient sensor streams, temporal feature extraction is performed using a modality-specific convolutional encoder. The encoder applies stacked Conv1D–BatchNorm–ReLU transformations with intermediate dropout to learn compact temporal representations from the input sequence, as defined in Equation 5.5. The resulting representation $H_m \in \mathbb{R}^{T \times D}$ preserves the temporal structure of the sensor sequence while projecting modality-specific features into a shared embedding space of dimension $D = 128$. As illustrated in Fig. 5.3, each modality is processed independently using three convolutional blocks, enabling the extraction of hierarchical temporal features before multimodal fusion. The convolutional layers progressively increase the number of feature channels ($n \rightarrow 64 \rightarrow 128 \rightarrow 128$), enabling the encoder to learn increasingly abstract temporal representations before multimodal fusion.

$$H_m = \phi_m(\mathbf{X}_m) \quad (5.5)$$

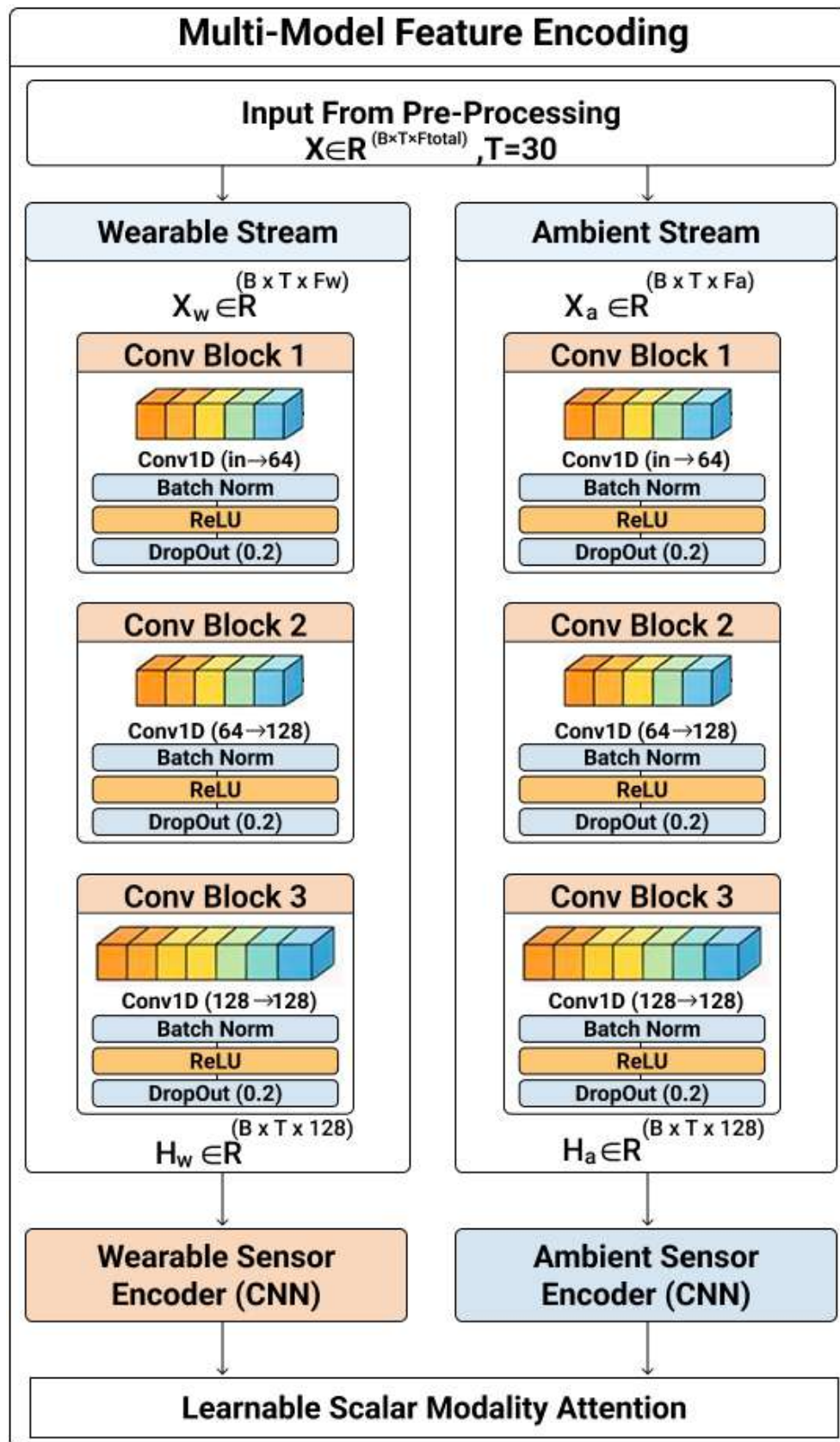


Figure 5.3: Modality-specific convolutional encoders for wearable and ambient sensor streams with temporal feature extraction

5.2.3 Learnable Attention-Based Fusion

Following modality-specific encoding, the wearable and ambient feature representations are integrated using a learnable scalar attention-based fusion mechanism. A trainable parameter vector $\mathbf{a} \in \mathbb{R}^2$ assigns relative importance to each modality, and the corresponding modality weights are normalized using the softmax operation defined in Equation 5.6. Unlike query-key-value attention mechanisms, this formulation performs global modality-level weighting, where the learned importance values are applied uniformly across all temporal steps.

$$\alpha = \text{softmax}(\mathbf{a}), \quad \alpha = [\alpha_w, \alpha_a], \quad \alpha_w + \alpha_a = 1 \quad (5.6)$$

The modality-specific feature representations are then scaled according to the learned attention weights, as expressed in Equation 5.7.

$$\tilde{H}_w = \alpha_w H_w, \quad \tilde{H}_a = \alpha_a H_a \quad (5.7)$$

The weighted representations are subsequently concatenated along the feature dimension and projected through the fusion transformation defined in Equation 5.8, where $\phi_f(\cdot)$ consists of a linear layer followed by Layer Normalization, ReLU activation, and Dropout.

$$H_f = \phi_f([\tilde{H}_w; \tilde{H}_a]) \quad (5.8)$$

As illustrated in Fig. 5.4, modality-specific features are weighted using learnable scalar attention and combined through a linear fusion block to obtain a unified representation. The resulting fused sequence $H_f \in \mathbb{R}^{T \times D}$ is forwarded to the transformer encoder for temporal modeling. The resulting fused sequence $H_f \in \mathbb{R}^{T \times D}$ is forwarded to the transformer encoder for temporal modeling. The learnable weighting mechanism enables adaptive balancing of wearable and ambient sensor contributions during training.

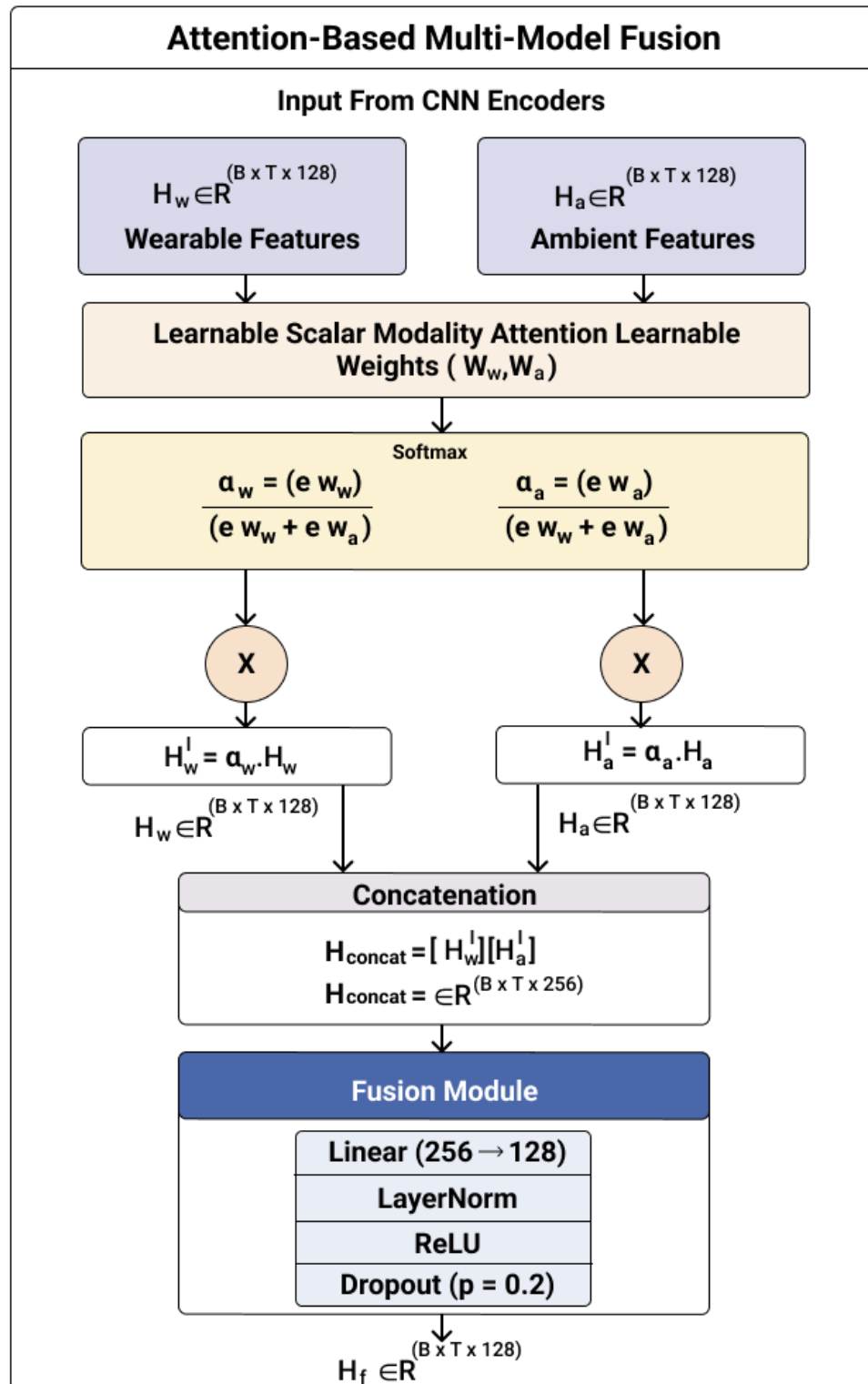


Figure 5.4: Attention-based scalar fusion of wearable and ambient features using learnable modality weights and fusion projection

5.2.4 Transformer Encoder for Temporal Modeling

The fused feature sequence $H_f \in \mathbb{R}^{T \times D}$ is processed by a stack of $L = 3$ transformer encoder layers to model long-range temporal dependencies within each sensor window. The transformer encoder operates directly on H_f without explicit positional encoding. The convolutional encoders preceding the transformer reduce redundancy in the raw sensor inputs and project them into a compact embedding space, which helps mitigate the quadratic computational complexity of self-attention with respect to sequence length [66].

Within each transformer encoder layer, multi-head self-attention is applied to model temporal relationships within the fused feature sequence. Linear projections are first used to generate query, key, and value representations from the fused embedding sequence, as defined in Equation 5.9.

$$Q = H_f W_Q, \quad K = H_f W_K, \quad V = H_f W_V \quad (5.9)$$

The attention output is then computed using scaled dot-product attention according to Equation 5.10, with dropout applied to the attention weights during training.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (5.10)$$

Each encoder layer follows a post-normalization structure with residual connections. A position-wise feed-forward network further refines the learned temporal representations, as expressed in Equation 5.11.

$$\text{FFN}(x) = W_2 \text{ReLU}(W_1 x) \quad (5.11)$$

After transformer encoding, adaptive average pooling is applied along the temporal dimension to obtain a fixed-length global representation, as defined in Equation 5.12.

$$\mathbf{z} = \text{AvgPool}\left(H_{\text{enc}}^f\right), \quad \mathbf{z} \in \mathbb{R}^D \quad (5.12)$$

The self-attention mechanism enables the model to capture dependencies between

any two time steps within the input sequence irrespective of their temporal distance, which is important for recognizing activities with complex temporal dynamics.

5.3 Hierarchical Multi-Task Learning

The shared pooled representation $\mathbf{z} \in \mathbb{R}^D$ is forwarded to two parallel task-specific classification heads for simultaneous prediction of locomotion primitives and high-level daily activities. Each classification head consists of a fully connected layer followed by ReLU activation and dropout, and a final linear projection to the corresponding class space. The task-specific output formulation is defined in Equation 5.13, where $k \in \{\text{loc}, \text{hl}\}$ denotes the locomotion and high-level activity prediction branches, respectively. The final predictions for both tasks are obtained independently from the corresponding task-specific outputs generated from the shared representation $\mathbf{z}$.

$$o^k = W_k^{(2)} \sigma \left(W_k^{(1)} \mathbf{z} \right), \quad k \in \{\text{loc}, \text{hl}\} \quad (5.13)$$

5.3.1 Locomotion-Level Activities

The locomotion-level task focuses on recognizing primitive physical activities such as standing, walking, sitting, and lying, which are primarily characterized by body motion patterns captured through wearable sensor signals. Learning these low-level activities enables the framework to model fundamental motion dynamics present in the sensor data, which further contributes to improving the shared representation used for joint multi-task learning.

5.3.2 High-Level Daily Activities

The high-level activity task focuses on recognizing complex daily activities such as preparing food, cleaning, and relaxing, which depend more strongly on contextual and environmental information captured through ambient sensors. Unlike locomotion primitives that are primarily motion-driven, high-level activities require the integration of object interaction and environmental context for reliable recognition. In the MM-HTT framework, the term "hierarchical" refers to the semantic relationship between locomotion primitives and high-level daily activities rather than a directed dependency between the two prediction heads. Both tasks are predicted independently from the shared temporal representation $\mathbf{z}$ using the task-specific formulation defined in Equation 5.13. This design enables multi-task regularization, where the shared representation is jointly

optimized to capture both low-level motion dynamics and high-level contextual activity patterns.

5.4 Experimental Setup

This section presents the experimental configuration used to evaluate the proposed multimodal framework under cross-user settings. The evaluation is conducted on three publicly available datasets using a unified preprocessing pipeline, consistent training configuration, and a subject-independent evaluation protocol. In addition, modality-specific experiments are conducted to analyze the contribution of wearable and ambient sensing within the framework.

5.4.1 Datasets: Opportunity++, PAMAP2, DSADS

The MM-HTT framework is evaluated on three publicly available HAR benchmark datasets offering diverse sensing configurations and activity scenarios. Table 5.1 summarizes the characteristics of the datasets used for evaluating the MM-HTT framework.

Opportunity++ [40, 60]. The Opportunity++ dataset contains synchronized recordings from wearable inertial sensors and ambient sensors deployed in a kitchen environment, with annotations for both locomotion primitives and high-level daily activities. Data were collected from four subjects performing six runs each, including five daily living activity runs and one drill run. This dataset is the primary benchmark for the multi-task learning and multimodal fusion evaluation, as it uniquely provides both wearable and ambient sensor modalities alongside dual-level activity annotations. The dataset is significantly class-imbalanced due to the dominance of the Null (non-activity) class, which is addressed through macro-averaged metric evaluation and Null-class exclusion from loss computation.

PAMAP2 [41]. The PAMAP2 dataset provides wearable inertial sensor recordings from three body locations (hand, chest, ankle) for nine subjects performing 18 physical activities including walking, running, cycling, and household tasks. It provides single-level activity annotations and does not include ambient sensing modalities; for this dataset, the framework is evaluated as a wearable-only multi-sensor fusion model.

DSADS [42]. The Daily and Sports Activities Dataset contains recordings from eight subjects performing 19 activities using inertial sensors placed on five body segments. Similar to PAMAP2, it provides single-level annotations without ambient sensing.

Table 5.1: Summary of datasets used for MM-HTT evaluation

Dataset	Subj.	Acts.	Modality	Annotation	Window
Opportunity++	4	Loco (5) + HL (6)	Wearable + Ambient	Dual-level	$T = 30, S = 15$
PAMAP2	9	18	Wearable (3 locs)	Single-level	$T = 30, S = 15$
DSADS	8	19	Wearable (5 locs)	Single-level	$T = 50, S = 25$

Continuous multivariate sensor streams were standardized using z-score normalization applied independently to each feature dimension. To prevent data leakage under the LOSO evaluation protocol, normalization parameters were computed exclusively from the training subjects. Wearable sensors provide dense continuous inertial signals, whereas ambient sensors generate sparse event-driven observations. After temporal alignment, all sensor streams were segmented into fixed-length overlapping windows. For a sequence $\{\mathbf{x}_1, \dots, \mathbf{x}_L\}$, each window is defined according to Equation 5.14, where T denotes the window length and S represents the stride between consecutive windows.

$$\mathbf{W}_i = \{\mathbf{x}_{iS+1}, \mathbf{x}_{iS+2}, \dots, \mathbf{x}_{iS+T}\} \quad (5.14)$$

The segmented dataset is subsequently represented as a three-dimensional tensor $\mathbf{X} \in \mathbb{R}^{N \times T \times F}$, as defined in Equation 5.15, where N denotes the number of windows, T the temporal length, and F the number of sensor features.

$$\mathbf{X} \in \mathbb{R}^{N \times T \times F} \quad (5.15)$$

For multimodal datasets, modality-specific feature tensors are extracted from the full sensor representation using the feature index set corresponding to each modality, as expressed in Equation 5.16.

$$\mathbf{X}_m = \mathbf{X}[:, :, \tilde{S}_m] \quad (5.16)$$

where $\tilde{S}_m$ denotes the feature indices associated with modality m . For the Opportunity++ dataset, each window is assigned dual labels $(y_{\text{loc}}, y_{\text{hl}})$ corresponding to locomotion and high-level activities, whereas PAMAP2 and DSADS use single-label activity annotations. Samples corresponding to the Null class are excluded from loss computation and evaluation to reduce bias caused by class imbalance.

5.4.2 Training Configuration and Implementation Details

The proposed MM-HTT framework is trained using supervised mini-batch optimization with batch size $B = 64$. Model parameters are optimized using the AdamW optimizer with learning rate 5×10^{-4} and weight decay 10^{-4} . Gradient clipping with maximum ℓ_2 -norm of 1.0 is applied during backpropagation to stabilize optimization during transformer training. A ReduceLROnPlateau learning rate scheduler is employed with reduction factor 0.5 and patience of 5 epochs to reduce the learning rate when validation performance saturates. The framework is trained for a maximum of 10 epochs, and the model checkpoint corresponding to the minimum validation loss is retained for evaluation. Dropout regularization is applied within convolutional encoders, fusion layers, and transformer encoder blocks to reduce overfitting under subject-independent evaluation settings.

During training, sensor windows are forwarded through modality-specific convolutional encoders, the scalar attention-based fusion module, transformer-based temporal encoder, and task-specific prediction heads. For the Opportunity++ dataset, a weighted multi-task objective is used to jointly optimize locomotion and high-level activity recognition tasks with task weights $w_{\text{loc}} = 1.0$ and $w_{\text{hl}} = 0.8$. Each task loss is computed using categorical cross-entropy with label smoothing ($\epsilon = 0.1$), while samples corresponding to the Null class are excluded from loss computation using an ignore-index mechanism. For PAMAP2 and DSADS, which contain single-level activity annotations, standard single-task categorical cross-entropy loss is employed. Validation performance is monitored after each epoch using the validation subset derived exclusively from training subjects under the LOSO protocol. The complete training procedure followed for the MM-HTT framework is presented in Algorithm 1.

Algorithm 1 Training procedure for the proposed MM-HTT framework

Require: Training dataset $\mathcal{D}$, validation dataset $\mathcal{V}$

```

1: Initialize model parameters  $\theta$ 
2: Initialize AdamW optimizer (lr= $5 \times 10^{-4}$ , weight decay= $10^{-4}$ )
3: Initialize ReduceLROnPlateau scheduler (factor=0.5, patience=5)
4: Set best validation loss =  $\infty$ 
5: Set patience counter = 0
6: for each epoch do
7:   for each mini-batch  $(\mathbf{X}, \mathbf{y})$  in  $\mathcal{D}$  do
8:     Compute predictions  $\hat{\mathbf{y}} = f_{\theta}(\mathbf{X})$ 
9:     if multi-task labels are available then
10:      Compute weighted multi-task loss (excluding Null samples using ignore-index)
11:    else
12:      Compute single-task cross-entropy loss
13:    end if
14:    Backpropagate gradients
15:    Clip gradients (max norm = 1.0)
16:    Update parameters using AdamW
17:  end for
18:  Compute validation loss on  $\mathcal{V}$ 
19:  Update scheduler using validation loss
20:  if validation loss improves then
21:    Save model checkpoint
22:    Reset patience counter
23:  else
24:    Increment patience counter
25:  end if
26:  if patience counter  $\geq 15$  then
27:    Stop training
28:  end if
29: end for
30: return Best-performing model

```

5.4.3 LOSO Evaluation Protocol

Cross-user generalization is evaluated using the LOSO cross-validation protocol. Let the dataset be represented as $\mathcal{D} = \{S_1, S_2, \dots, S_N\}$, where each S_i denotes the data corresponding to subject i , as defined in Equation 5.17. For each LOSO fold, one subject is held out for testing while the remaining subjects are used for training according to the partitioning strategy defined in Equation 5.18. The training portion is further divided into 90% training and 10% validation subsets using only data from the training subjects, ensuring that no information from the held-out test subject is used during model development. This evaluation protocol results in four LOSO folds for Opportunity++, eight folds for DSADS, and nine folds for PAMAP2. Performance metrics are computed on the held-out subject for each fold and averaged across all folds to obtain the final reported results.

$$\mathcal{D} = \{S_1, S_2, \dots, S_N\} \quad (5.17)$$

$$\mathcal{D}_{\text{train}} = \mathcal{D} \setminus S_i, \quad \mathcal{D}_{\text{test}} = S_i \quad (5.18)$$

Due to large class imbalances in the Opportunity++ dataset, the macro F1-score is chosen as the main evaluation metric to provide a balanced assessment of all activity classes [83]. To avoid bias towards the non-activity samples that dominate the total number of samples, the samples of the Null class are excluded from the aggregate metric calculations. Evaluations for Opportunity++ are completed separately for locomotion vs. high-level activity recognition tasks, whereas PAMAP2 and DSADS are evaluated using a single activity recognition task. The evaluations assess how well multimodal fusion and multi-task learning strategies work within the framework by evaluating the respective modality configurations using the same experimental conditions to ensure a fair comparison.

5.5 Lightweight Model for Edge Deployment

A lightweight version of the framework has been constructed for use in resource-limited settings e.g., edge devices and wearable devices to allow deployment. The low resource model is an alternative to the complete MM-HTT framework and supports dual task prediction while providing a high performance computationally. Unlike the full architecture, the lightweight model does not employ modality-specific encoders, attention-based

fusion, or transformer-based temporal modeling. Instead, the lightweight model operates on the concatenated sensor window $\mathbf{X} \in \mathbb{R}^{T \times F}$ and performs temporal feature extraction using depthwise separable one-dimensional convolutions, which decompose standard convolutions into depthwise and pointwise operations to reduce computational complexity [84]:

$$\mathbf{z}_S = g_\phi(\mathbf{X}) \quad (5.19)$$

where $g_\phi(\cdot)$ denotes the depthwise separable convolutional feature extractor. The extracted features are aggregated using adaptive average pooling and forwarded to task-specific linear classification heads, preserving the multi-task prediction structure. The lightweight model is trained independently using the same supervised learning procedure as the full framework and does not employ knowledge distillation. Table 5.2 presents a comparison between the full MM-HTT framework and the lightweight model in terms of parameter count and model size.

Table 5.2: Efficiency comparison between the full MM-HTT framework and the lightweight model

Model	Parameters	Model Size (MB)
MM-HTT (Full Model)	1.16M	4.48
Lightweight Model	0.009M	0.05
Reduction	$\approx 122\times$	$\approx 89\times$

As observed in Table 5.2, the lightweight model reduces the number of parameters from 1.16M to 0.009M and the model size from 4.48 MB to 0.05 MB, corresponding to approximately $122\times$ and $89\times$ reductions, respectively. These reductions improve computational efficiency and make the model suitable for deployment in resource-constrained environments. While the lightweight architecture provides significant efficiency gains, detailed evaluation of performance trade-offs and hardware-level metrics such as inference latency is not conducted in this study and remains an area for future investigation.

5.6 Results and Analysis

This section presents the experimental evaluation of the proposed framework on the Opportunity++, PAMAP2, and DSADS datasets under the LOSO protocol. Performance is evaluated using accuracy, macro-averaged precision, recall, macro F1-score, and weighted F1-score, with macro F1-score used as the primary metric due to class imbalance in the Opportunity++ dataset.

5.6.1 Overall Performance

The multimodal fusion configuration achieves the highest performance across both locomotion and high-level activity recognition tasks. Locomotion recognition reaches a macro F1-score of 0.9290, while high-level activity recognition achieves 0.8111, indicating that combining motion-related wearable sensing with contextual ambient information improves recognition performance across both activity levels. The wearable-only configuration performs strongly for locomotion recognition due to its ability to capture periodic body motion patterns, whereas ambient sensing contributes more effectively to high-level activity recognition through contextual information. The fusion configuration consistently outperforms both individual modalities, demonstrating the effectiveness of the proposed scalar attention-based fusion mechanism in integrating complementary sensor information. Table 5.3 presents the modality ablation results on the Opportunity++ dataset. Figure 5.5 further illustrates the class-wise prediction behavior through confusion matrices and one-vs-rest ROC curves for the fusion configuration, showing improved discrimination across both locomotion and high-level activity classes.

Table 5.3: Sensor modality ablation results on the Opportunity++ dataset

Setting	Task	Acc (%)	Prec (M)	Rec (M)	F1 (M)	F1 (W)
Fusion (All)	Locomotion	91.22	0.9269	0.9313	0.9290	0.9120
	High-Level	81.49	0.8186	0.8049	0.8111	0.8146
Ambient Only	Locomotion	74.36	0.7663	0.7660	0.7647	0.7389
	High-Level	78.60	0.7818	0.7969	0.7870	0.7867
Wearable Only	Locomotion	90.53	0.9235	0.9240	0.9232	0.9046
	High-Level	73.73	0.7385	0.7013	0.7037	0.7206

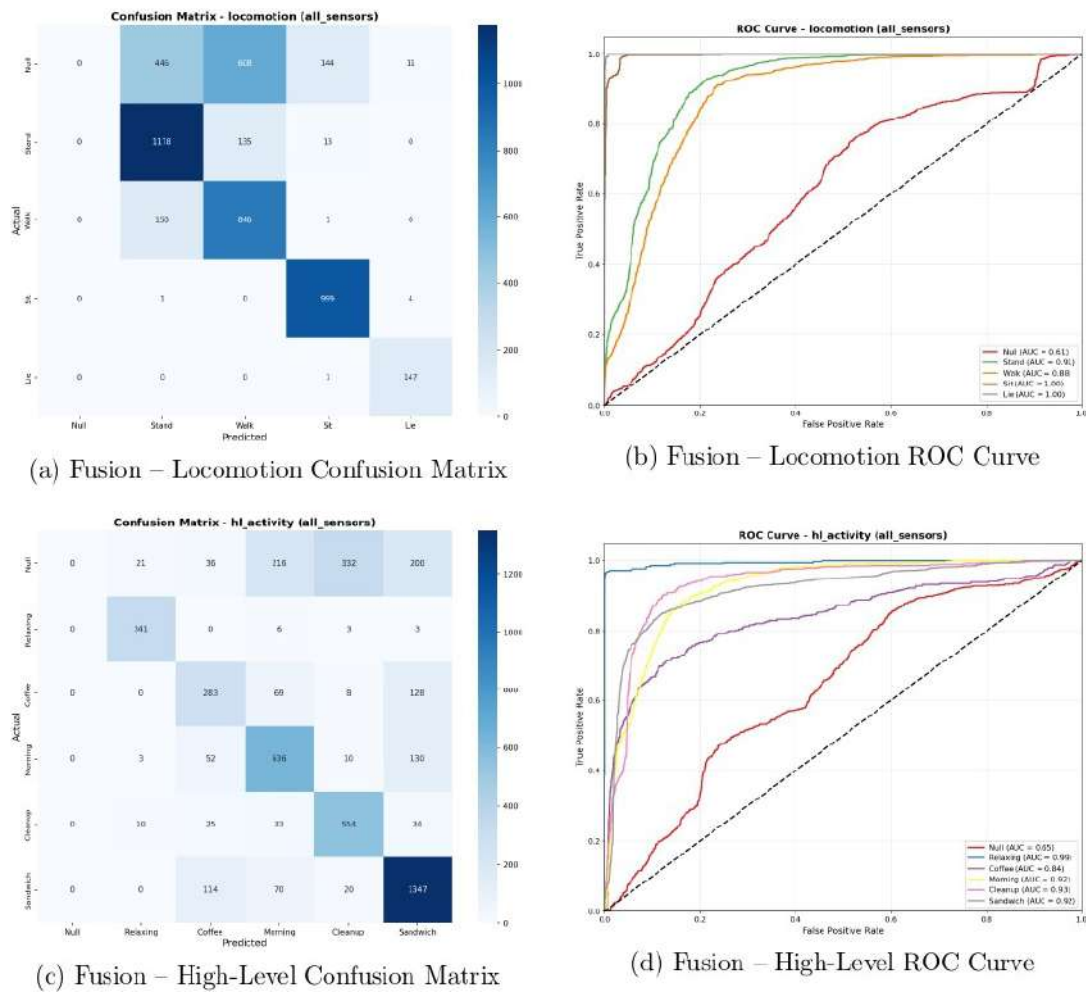


Figure 5.5: Confusion matrix and one-vs-rest ROC curves for the fusion configuration on the Opportunity++ dataset

Table 5.4: Sensor ablation results on DSADS and PAMAP2 datasets

Dataset	Setting	Acc (%)	Prec (M)	Rec (M)	F1 (M)	F1 (W)
DSADS	Arms Only	64.18	0.6405	0.6418	0.6045	0.6045
	Torso Only	75.09	0.7618	0.7509	0.7425	0.7425
	Legs Only	83.67	0.8988	0.8367	0.8405	0.8405
	Fusion (All)	84.81	0.8987	0.8481	0.8454	0.8454
PAMAP2	Chest Only	61.98	0.5881	0.5530	0.5545	0.6335
	Ankle Only	63.15	0.6524	0.6585	0.6210	0.6297
	Hand Only	89.31	0.8953	0.8830	0.8825	0.8912
	Fusion (All)	92.03	0.9270	0.9211	0.9183	0.9199

Across both datasets, multimodal fusion consistently achieves the highest performance, demonstrating the benefit of combining complementary sensor locations.

5.6.2 Cross-User Generalization

Table 5.5 presents the LOSO cross-user evaluation results across all datasets. The framework demonstrates consistent cross-user generalization across datasets. The performance of locomotion recognition is consistently stable because of the regularity of motions, but the performance of high-level activity recognition is much more variable due to its dependence on context. Figure 5.6 shows examples of the performance of different classifiers on the DSADS and PAMAP2 datasets via confusion matrices and ROC curves and thus further establishes that the framework works well across users.

Table 5.5: LOSO cross-user evaluation results

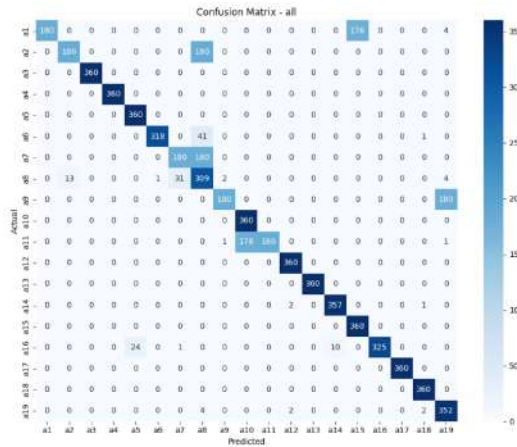
Dataset	Task	Acc (%)	Prec (M)	Rec (M)	F1 (M)	F1 (W)
Opportunity++	Locomotion	85.14	0.8585	0.8164	0.8021	0.8473
	High-Level	72.61	0.7417	0.7220	0.7183	0.7236
DSADS	Average	87.40	0.8629	0.8740	0.8492	0.8492
PAMAP2	Average	87.61	0.8717	0.8369	0.8261	0.8696

5.6.3 Effect of Multimodal Fusion and Temporal Modeling

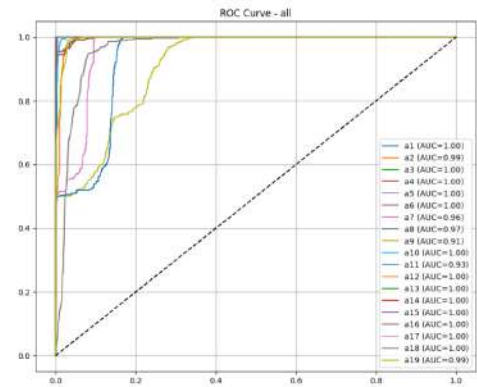
Multimodal fusion consistently demonstrates better performance in recognizing activities than single-mode configurations across every tested dataset. Movement sensing through wearables would contribute strongly towards recognizing locomotion, while context information through ambient sensors would contribute further positively towards higher-level activity recognition than other methods. The proposed fusion mechanism effectively combines these complementary modalities, while transformer-based temporal modeling captures long-range dependencies within activity sequences.

5.7 Discussion

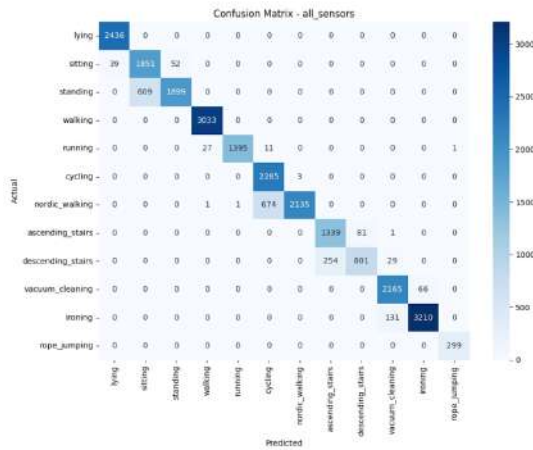
The experimental results demonstrate that multimodal fusion and transformer-based temporal modeling improve cross-user activity recognition across diverse sensing environments. The framework maintains stable locomotion recognition under LOSO evaluation, while high-level activities remain comparatively challenging due to con-



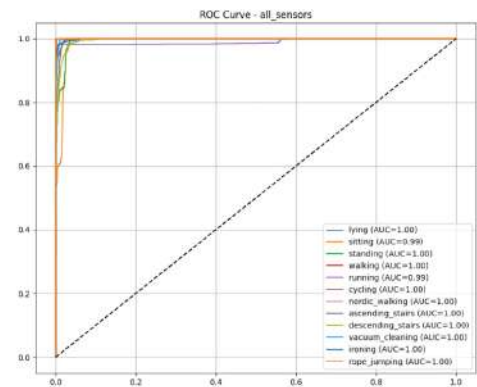
(a) DSADS – Confusion Matrix



(b) DSADS – ROC Curve



(c) PAMAP2 – Confusion Matrix



(d) PAMAP2 – ROC Curve

Figure 5.6: Confusion matrices and ROC curves for DSADS and PAMAP2 datasets under the fusion configuration

textual and inter-subject variability [5]. The lightweight variant further shows that substantial reductions in parameter count and model size are achievable for resource-constrained deployment, although with reduced representational capacity compared to the full transformer-based architecture. Future work may investigate knowledge distillation strategies for transferring temporal representations from the full MM-HTT framework to lightweight models [85].

5.8 Summary

This chapter presented the MM-HTT framework for multimodal HAR using synchronized wearable and ambient sensor streams under cross-user evaluation settings. The framework integrates modality-specific convolutional encoders, scalar attention-based multimodal fusion, transformer-based temporal modeling, and multi-task learning for

simultaneous locomotion and high-level activity recognition. Evaluation on the Opportunity++, PAMAP2, and DSADS datasets under the LOSO protocol demonstrated the effectiveness of multimodal fusion for improving recognition performance and cross-user generalization. A lightweight architectural variant based on depthwise separable convolution was also investigated to reduce computational complexity for resource-constrained deployment scenarios.

CHAPTER 6

RESULTS AND COMPARATIVE DISCUSSION

This chapter presents a comparative analysis of the three research works by examining differences in input representation, model design, performance characteristics, and deployment feasibility. The focus of the discussion is on identifying practical trade-offs and on developing the understanding of the relationship between the approaches in terms of accuracy, robustness and applicability in the real world.

6.1 Comparative Analysis Across the Three Studies

The approaches vary from each other mainly in the input representation and the abstraction level of the activity recognition. These three approaches are based on vision, skeleton, and sensor, respectively, which respectively leverage image, skeleton-based, and sensor-based representations of the structure of the human body. Table 6.1 provides a cross comparison between these approaches with respect to input representation, model design and characteristics.

The progression in the comparison in Table 6.1 shows how image-based input is taken and how it is passed on to more structured representations of keypoints and then onto multimodal sensor data. The change from images to keypoints diminishes the dependency of background and enhances the stability of the recognition. The integration of sensor data further improves robustness, especially when varying from one subject to another and under changing environmental conditions. Vision-based approaches allow light-weight models and have the advantage of providing real-time performance, but they are susceptible to occlusion and illumination changes. These skeleton-based approaches can be used to represent spatial data in a structured way to improve the robustness of the approach, while simultaneously being computationally efficient. In addition to providing better generalization and privacy, sensor-based approaches introduce some system complexity without relying on visual input. The observations here show that the gap between the approaches is due to different trade-offs between accuracy, computational demands, and deployment issues rather than any

single best approach.

Table 6.1: Comparison across the three approaches based on input representation, model design, and performance characteristics

Dimension	Pose Estimation	Skeleton-based HAR	Multimodal HAR
Input	Images / video frames	Skeleton keypoints	Wearable and ambient sensors
Model	CNN and Transformer	CNN-LSTM	CNN encoder with Transformer
Task	Keypoint localization	Activity classification	Multi-task activity recognition
Accuracy	Up to 80.6 mAP	Up to 94%	Up to 92.03%
Latency	1–10 ms (lightweight)	125 ms (CPU)	Not evaluated on hardware
Real-time capability	Supported	Supported	Moderate
Generalization	Limited	Moderate	High
Modality	Vision	Vision (skeleton)	Multimodal
Edge deployment	Supported	Supported	Supported (lightweight variant)
Privacy	Low	Low	High
Hierarchy	Not considered	Not considered	Locomotion and high-level activities
Evaluation	Benchmark datasets	Train/validation/test split	LOSO cross-validation

6.2 Vision-based and Sensor-based Activity Recognition: Trade-offs

Vision-based and sensor-based approaches vary in terms of input modality and environmental dependence, and deployment requirements. In contrast, vision-based methods are based on image or video data, while being based on wearable and ambient signals,

sensor-based methods capture motion patterns. In vision-based approaches, detailed spatial representation of human motions is provided and activity interpretation is enabled by the body structure itself. Occlusion, illumination and the viewpoint of the camera have an impact on the accuracy of the pose estimate and subsequent recognition. Skeleton-based representations decrease the need for background information through the use of keypoint-based features derived from pose estimation as part of their processing, as they are more robust than image-based methods without needing to sacrifice efficiency for real-time processing.

Sensor-based methods don't rely on visual data and can't be impacted by occlusion or lighting changes. Motion signals are captured by wearable and ambient sensors, and are consistent across environments, enhancing privacy since no visual information is needed. These approaches require sensor placement and synchronization, however, and there is no explicit spatial structure. Sensor-based approaches tested in cross-subject environments demonstrate much greater user-independence, whereas vision-based methods are sensitive to environmental changes because they rely on visual information. The different approaches shown in Figure 6.1 are a compromise between spatial representation, robustness to the environment, and deployment requirements.

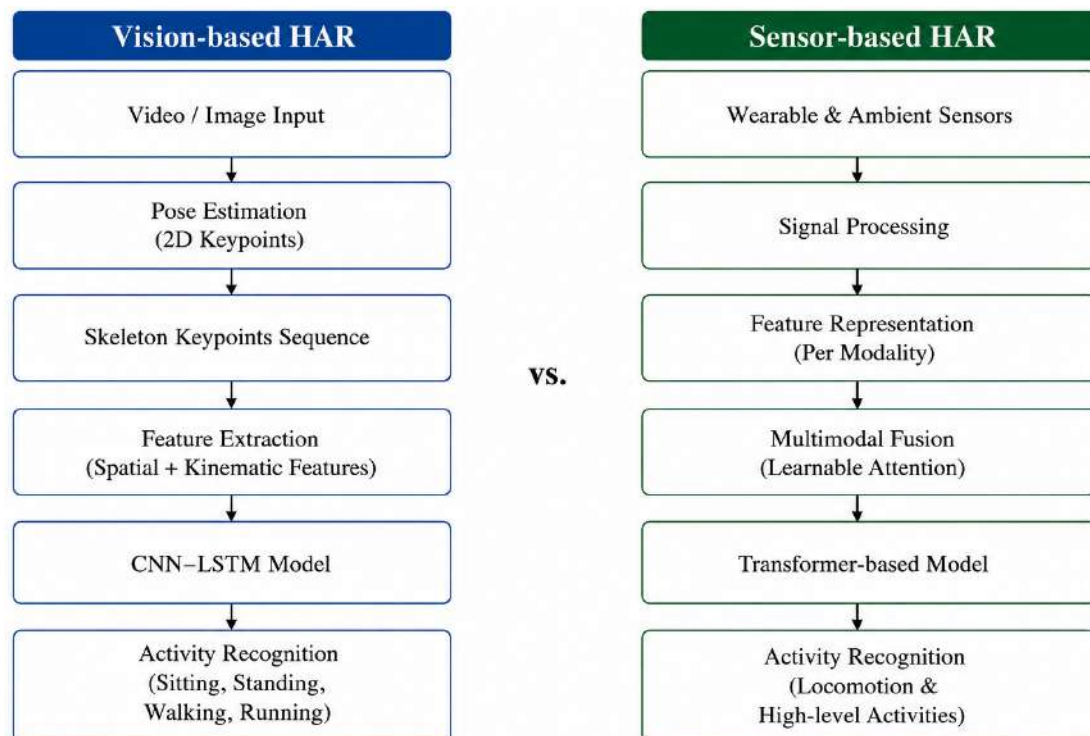


Figure 6.1: Comparison between vision-based and sensor-based HAR approaches

6.3 Accuracy, Latency, and Deployment Feasibility

The comparison of the three approaches suggests a compromise between recognition scores, computational complexity, and implementation potential. The CNN-based pose estimation models are lightweight and are suitable for real-time and edge-based applications as they offer low inference latency, while transformer-based models result in higher recognition performance with higher computational and memory consumption. The performance of the recognition and the relative computational complexity of representative pose estimation, skeleton-based HAR and multimodal HAR are shown in Figure 6.2. Comparisons of pose estimation models reveal the transition from lightweight models like OpenPose and MoveNet for CNN to heavier models like ViTPose for transformer. It enhances the representational power and the more accurate the result, the more expensive it is to compute and the more complex the model. For low latency inference and deployment on resource-limited platforms, lightweight models continue to be the preferred option.

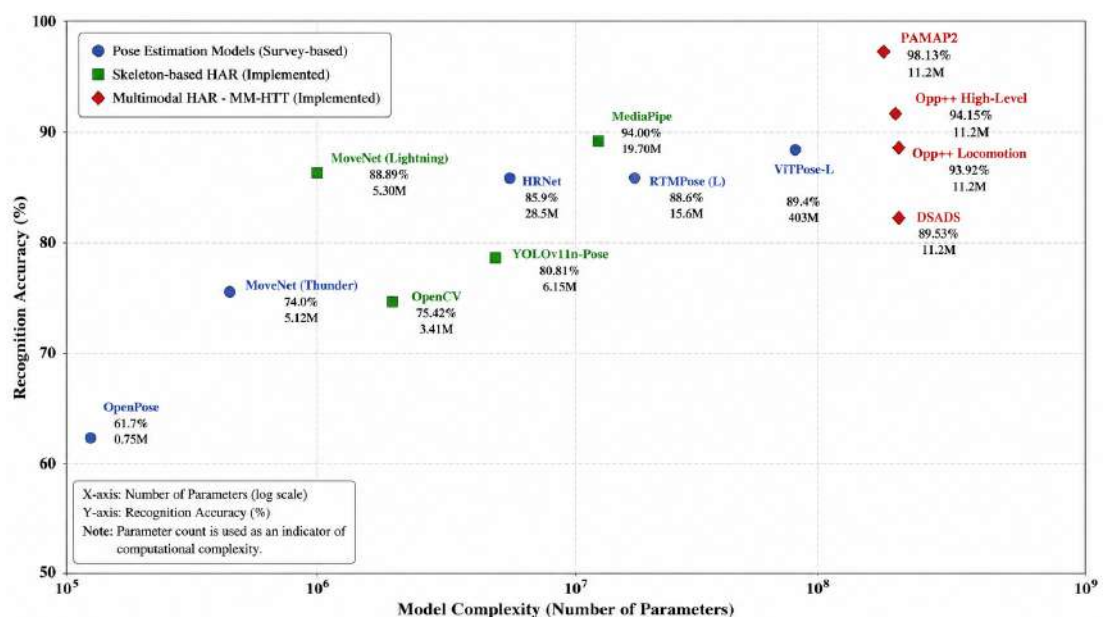


Figure 6.2: Comparison of recognition performance and model complexity across the analyzed approaches

Skeleton-based activity recognition provides a balance between recognition performance and computational efficiency by utilizing compact keypoint representations derived from pose estimation. Among the evaluated backends, MediaPipe achieved the highest recognition accuracy, whereas YOLOv11 Pose achieved lower latency and smaller model size, supporting real-time deployment. Multimodal sensor-based ap-

proaches improve robustness through the integration of wearable and ambient sensor information. The MM-HTT framework demonstrated consistent performance across multiple datasets under the LOSO evaluation protocol. The lightweight MM-HTT variant further reduced parameter count and model size, supporting deployment in resource-constrained environments. However, hardware-level latency evaluation was not included. Deployment suitability depends on application requirements and operational constraints. Vision-based approaches are appropriate when spatial interpretation and camera-based monitoring are required. Skeleton-based methods reduce background dependency while maintaining real-time feasibility. Sensor-based approaches are more suitable for privacy-sensitive environments and scenarios requiring improved cross-user robustness. Table 6.2 summarizes the deployment characteristics of the three approaches.

Table 6.2: Deployment characteristics across the analyzed approaches

Requirement		Pose Estimation	Skeleton-based HAR	Multimodal HAR
Camera requirement	require-	Required	Required	Not required
Privacy sensitivity		Limited	Limited	High
Real-time capability	capabil-	Supported	Supported	Moderate
Cross-user generalization	general-	Limited	Moderate	High
Complex activity recognition	activity	Limited	Moderate	Supported
Edge deployment		Supported	Supported	Supported (lightweight variant)
Hierarchical activity modeling	activ-	Not considered	Not considered	Supported
Deployment complexity	com-	Low	Moderate	Moderate to high

6.4 Pipeline from Pose Estimation to Sensor-based Activity Recognition

The three approaches represent different stages in the progression of HAR systems. Pose estimation models operate on image or video data to extract body keypoints, which are then used for skeleton-based activity recognition. Sensor-based approaches further extend activity recognition by incorporating wearable and ambient sensor information. The progression shown in Figure 6.3 reflects the transition from visual body localization to activity recognition using structured skeletal representations and multimodal sensor information.

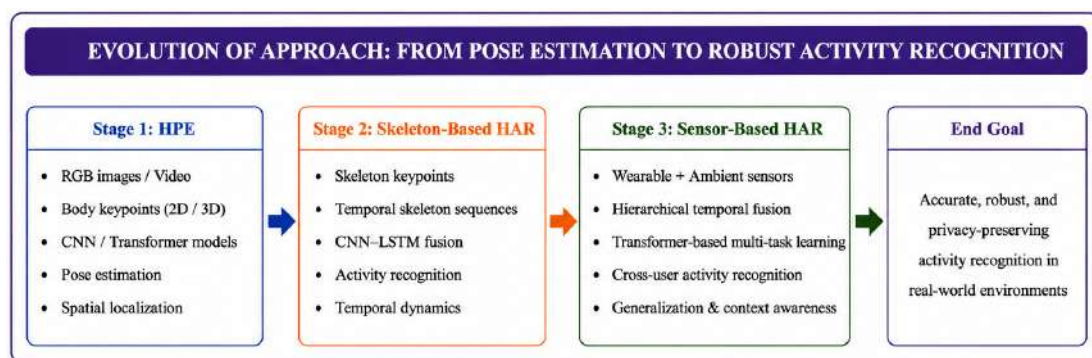


Figure 6.3: Progression from HPE to multimodal HAR

Pose estimation provides spatial representations of the human body through joint localization. These keypoints form the basis for skeleton-based activity recognition, where temporal motion patterns are learned using sequential DL models. The skeleton-based approach reduces background dependency while preserving motion information required for activity classification. Sensor-based activity recognition extends this process by incorporating motion and contextual information from wearable and ambient sensors. Unlike vision-based methods, sensor-based approaches operate independently of visual conditions and improve robustness across different environments and users. The multimodal framework further combines information from multiple sensing modalities to support recognition of both locomotion and high-level activities.

6.5 Limitations

The limitations identified across the three studies are related to dataset diversity, deployment evaluation, and generalization across different environments and subjects. These limitations define scope of reported findings and indicate directions for further investigation.

Dataset diversity and evaluation scope. The activity recognition study using the skeleton was tested on a small number of activities, while in a controlled environment. The evaluation was conducted without support from multiple users, large differences in perspectives or occlusion-rich situations, which could impact performance in more complex real-world scenarios. The multimodal sensor-based framework was tested on publicly available data sets like Opportunity++, PAMAP2 and DSADS using LOSO validation, but these data sets have been acquired in specific experimental conditions and have a relatively small number of subjects. Similarly, the pose estimation analysis focused on benchmark datasets commonly used for model comparison, which may not fully represent unconstrained deployment environments.

Hardware-level deployment analysis. Although lightweight architectures and model compression strategies were explored, hardware-level deployment evaluation was not performed. The multimodal framework reported reductions in parameter count and model size for the lightweight variant, but inference latency, memory consumption, and power efficiency on embedded hardware platforms were not experimentally analyzed. In the skeleton-based study, latency measurements were obtained on desktop hardware and do not directly represent performance on low-power edge devices.

Cross-dataset generalization. No cross-dataset transfer experiments were performed and the evaluations were done independently within each dataset. So the strength of these learned representations in various sensing environments, activity distributions and acquisition protocols are yet to be verified. More evaluation on diverse data sets would give greater confidence in the ability to generalize.

Sensitivity to upstream representations and sensor configuration. Skeleton-based activity recognition relies on the accuracy of the pose estimation and keypoint detection. Joint localization might be inaccurate and influence the recognition of activity performance in later temporal modeling. Other sensor-based methods are also affected by the location, timing, and window segmentation of the sensors, all of which were set for experimentation.

Lightweight model characterization. The lightweight MM-HTT variant successfully reduced the model size and the number of parameters of the model with the same dual-task prediction structure. In contrast, there was no detailed quantitative comparison between the lightweight and full models on all datasets that was extensively analyzed. For real-world edge deployments, hardware-specific deployment analysis and efficiency benchmarking are still required. Table 6.3 summarizes the major limitations identified

across the three studies.

Table 6.3: Summary of limitations across the three studies

Limitation	HPE	Skeleton HAR	MM-HTT
Limited dataset diversity	✓	✓	Partial
No hardware-level deployment evaluation	✓	Partial	✓
No cross-dataset evaluation	✓	✓	✓
Sensitivity to pose estimation quality	N/A	✓	N/A
Sensitivity to sensor placement/configuration	N/A	N/A	✓
Lightweight model not extensively evaluated	N/A	N/A	✓

✓ = limitation present; Partial = partially addressed; N/A = not applicable

These limitations define the scope of the reported findings and indicate directions for further investigation discussed in Chapter 7.

6.6 Summary

This chapter presented a comparative discussion of the analyzed approaches by examining differences in input representation, model design, computational characteristics, and deployment feasibility. The discussion highlighted the trade-offs between vision-based, skeleton-based, and sensor-based activity recognition approaches in terms of robustness, privacy, real-time capability, and generalization. The progression from pose estimation to multimodal activity recognition was also discussed to establish the relationship between the three studies. In addition, the limitations associated with dataset diversity, deployment evaluation, and cross-dataset generalization were identified to define the scope of the reported findings.

CHAPTER 7

CONCLUSION, FUTURE WORK, AND SOCIAL IMPACT

This chapter presents the concluding remarks of the research work together with important future research directions and the broader social impact associated with DL-based HPE and HAR systems.

7.1 Conclusion

This research work investigated DL-based HPE and HAR using skeleton representations, spatial-temporal learning, and multimodal activity understanding. The evolution of pose estimation methods from lightweight real-time systems to transformer-based systems and discussed their practical applications in real-world activity recognition scenarios. The results showed that the lightweight CNN-based models, like MediaPipe and MoveNet, feature low computational complexity, small memory footprint, and enable efficient low-latency inference, making them ideal for edge-oriented and real-time applications, while the transformer-based models, capture global information about the image and model features based on attention, resulting in greater contextual understanding and robustness under challenging scenarios, such as occlusion, variation in viewing angle, and multi-person interactions. The skeleton-based activity recognition framework also showed that structured skeletal representations could effectively maintain the information of motion, and the motion information is not heavily reliant on the background appearance and environmental noise. The comparisons to MediaPipe Pose, MoveNet Thunder, OpenCV COCO and YOLOv11 Pose revealed key considerations such as trade-offs between recognition accuracy, inference latency, and deployment efficiency, with MediaPipe Pose showing superior classification accuracy and YOLOv11 Pose exhibiting lower latency for real-time deployment. The multimodal temporal learning framework added complementary motion and appearance information to boost the activity recognition performance with high robustness against variations of the subjects and environments. The transformer-based temporal modelling significantly enhanced the representation of long-range sequences and understanding of activity in the context of the sequence, especially in the cross-user evaluation scenario. The results show that

there is a need for a trade-off among all the factors affecting the recognition accuracy, temporal continuity, computational efficiency, scalability and deployment feasibility in various application environment in order to achieve an effective and powerful HAR system.

7.2 Future Work

The proposed approaches achieved good results in several evaluation scenarios, but a few directions are still open for further studies. With the goal of enhancing generalization capabilities in more realistic environments, future efforts could involve scaling up evaluation to larger and more varied datasets with crowded scenes, multiple-person interactions, complex transitional activities, extreme occlusion, viewpoint changes, and lighting changes. Further research on graph-based spatial-temporal learning, adaptive attention mechanisms, lightweight transformer architectures, and self-supervised learning could further enhance the robustness of temporal representation learning and recognition. Additionally, optimization methods like pruning, quantization, knowledge distillation, and hardware-aware Neural Architecture Design can be studied in the future to minimize the computational complexity without compromising recognition metrics, making it suitable for embedded systems, wearable devices, and edge-based intelligent environments. Further improvements in robustness can be achieved by combining the additional sensing modalities such as inertial, depth camera, thermal imaging, and physiological sensors, and more context-aware activity understanding systems could be developed.

7.3 Social Impact and Ethical Considerations

The presented approaches can be used in the fields of healthcare, rehabilitation, assisted living, intelligent surveillance, fitness analysis, and HCI systems. In this context, the social aspects and ethical concerns associated with intelligent activity recognition technologies, which are outlined below.

7.3.1 Healthcare and Rehabilitation

Real-time activity recognition systems can be used in posture assessment, fall detection, remote monitoring of physiotherapy, support in rehabilitation and continuous patient activity analysis. This can help make healthcare facilities more accessible and help people to continue to monitor without needing to be closely supervised at all times.

7.3.2 Elderly Care and Assistive Systems

Automated recognition of abnormal movement patterns and daily activities can enhance safety and help to support independent living through continuous activity monitoring. In smart home and assisted living settings, intelligent assistive systems can be used to further assist elderly people.

7.3.3 Fitness and HCI

Skeleton-based activity analysis can be used to assist in exercise monitoring, gesture interaction, movement feedback and immersive virtual interaction environments. These systems can be utilized for individualized fitness analysis and also for natural HCI applications.

7.3.4 Intelligent Surveillance and Workplace Safety

Under real-time operating conditions, activity recognition systems can help with anomaly detection, behavioral monitoring and safety analysis. These could be used to help intelligent surveillance and monitoring of workplaces in safety-critical applications.

7.3.5 Privacy and Ethical Concerns

Also, there is a concern with privacy, equity, responsible monitoring and ethical approaches to vision based monitoring, and sensor based systems. There can be differences in the recognitions of different subject groups and between different operating conditions as a result of bias in the data and environmental imbalance. The following are crucial challenges that need to be addressed in the future application of intelligent activity recognition systems: security of handling behavioral data, informed consent of the users, transparency in controlling the monitoring process and the fair evaluation. Assessment under diverse scenarios and populations will still be needed for safe and responsible adoption of DL based HAR technologies.

REFERENCES

- [1] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.
- [2] D. Bouchabou, S. M. Nguyen, C. Lohr, B. LeDuc, and I. Kanellos, "A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *Sensors*, vol. 21, no. 18, p. 6037, 2021.
- [3] A. Gupta, K. Gupta, K. Gupta, and K. Gupta, "A survey on human activity recognition and classification," in *2020 international conference on communication and signal processing (ICCSP)*, pp. 0915–0919, IEEE, 2020.
- [4] A. Hayat, F. Morgado-Dias, B. P. Bhuyan, and R. Tomar, "Human activity recognition for elderly people using machine and deep learning approaches," *Information*, vol. 13, no. 6, p. 275, 2022.
- [5] T. S. Qureshi, M. H. Shahid, A. A. Farhan, and S. Alamri, "A systematic literature review on human activity recognition using smart devices: advances, challenges, and future directions," *Artificial Intelligence Review*, vol. 58, no. 9, p. 276, 2025.
- [6] M. A. Hossen and P. E. Abas, "Machine learning for human activity recognition: State-of-the-art techniques and emerging trends," *Journal of Imaging*, vol. 11, no. 3, p. 91, 2025.
- [7] F. Serpush, M. B. Menhaj, B. Masoumi, and B. Karasfi, "Wearable sensor-based human activity recognition in the smart healthcare system," *Computational intelligence and neuroscience*, vol. 2022, no. 1, p. 1391906, 2022.
- [8] M. Alarfaj, A. Al Madini, A. Alsafran, M. Farag, S. Chtourou, A. Afifi, A. Ahmad, O. Al Rubayyi, A. Al Harbi, and M. Al Thunaian, "Wearable sensors based on artificial intelligence models for human activity recognition," *Frontiers in artificial intelligence*, vol. 7, p. 1424190, 2024.
- [9] S. Dill, A. Rösch, M. Rohr, G. Güney, L. De Witte, E. Schwartz, and C. Hoog Antink, "Accuracy evaluation of 3d pose estimation with mediapipe pose for physical exercises," in *Current Directions in Biomedical Engineering*, vol. 9, pp. 563–566, De Gruyter, 2023.
- [10] D. Ekambaram and V. Ponnusamy, "Real-time monitoring and assessment of rehabilitation exercises for low back pain through interactive dashboard pose analysis using streamlit—a pilot study," *Electronics*, vol. 13, no. 18, p. 3782, 2024.
- [11] F. Liang, Z. Su, and W. Sheng, "Multimodal monitoring of activities of daily living for elderly care," *IEEE Sensors Journal*, vol. 24, no. 7, pp. 11459–11471, 2024.

- [12] R. Singh and A. Sharma, "Stad-convbi-lstm: Spatio-temporal attention-based deep convolutional bi-lstm framework for abnormal activity recognition," *Journal of Visual Communication and Image Representation*, p. 104465, 2025.
- [13] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.
- [14] R. A. Hamad, W. L. Woo, B. Wei, and L. Yang, "Overview of human activity recognition using sensor data," in *UK Workshop on Computational Intelligence*, pp. 380–391, Springer, 2022.
- [15] A. Sharma and R. Singh, "Convst-lstm-net: Convolutional spatiotemporal lstm networks for skeleton-based human action recognition," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 34, 2023.
- [16] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, 2015.
- [17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [18] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th international conference on mobile computing, applications and services*, pp. 197–205, IEEE, 2014.
- [19] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.
- [20] G. AI, "Movenet: Ultra fast and accurate pose detection model." Online Resource, TensorFlow Blog, 2021. <https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html>.
- [21] X. Huang, Y. Yuan, C. Chang, Y. Gao, C. Zheng, and L. Yan, "Human activity recognition method based on edge computing-assisted and gru deep learning network," *Applied Sciences*, vol. 13, no. 16, p. 9059, 2023.
- [22] R. Abdel-Salam, R. Mostafa, and M. Hadhood, "Human activity recognition using wearable sensors: review, challenges, evaluation benchmark," in *International workshop on deep learning for human activity recognition*, pp. 1–15, Springer, 2021.
- [23] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

- [24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.
- [25] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8877–8886, 2023.
- [26] Y. Dai, Y. Lin, X. Lin, C. Wen, L. Xu, H. Yi, S. Shen, Y. Ma, and C. Wang, "Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 682–692, 2023.
- [27] G. Alam, I. McChesney, P. Nicholl, and J. Rafferty, "Open datasets in human activity recognition research—issues and challenges: A review," *IEEE Sensors Journal*, vol. 23, no. 22, pp. 26952–26980, 2023.
- [28] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.
- [29] L. Zhou, X. Meng, Z. Liu, M. Wu, Z. Gao, and P. Wang, "Human pose-based estimation, tracking and action recognition with deep learning: A survey," *arXiv preprint arXiv:2310.13039*, 2023.
- [30] J.-L. Chung, L.-Y. Ong, and M.-C. Leow, "Comparative analysis of skeleton-based human pose estimation," *Future internet*, vol. 14, no. 12, p. 380, 2022.
- [31] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2637–2646, 2022.
- [32] D. Debalaxmi, D. K. Vishwakarma, and V. Ranga, "Analyzing yoga pose recognition: A comparison of mediapipe and yolo keypoint detection with ensemble techniques," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 1011–1017, IEEE, 2024.
- [33] R. Singh and A. Sharma, "Occluded skeleton-based multi-stream model using part-aware spatial-temporal graph convolutional network for human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 156, p. 111183, 2025.
- [34] S. Xaviar, X. Yang, and O. Ardakanian, "Robust multimodal fusion for human activity recognition," *arXiv preprint arXiv:2303.04636*, 2023.
- [35] S. U. Khan, M. Sultana, S. Danish, D. Gupta, N. S. Alghamdi, S. Woo, D.-G. Lee, and S. Ahn, "Multimodal feature fusion for human activity recognition using human centric temporal transformer," *Engineering Applications of Artificial Intelligence*, vol. 160, p. 111844, 2025.

- [36] M. Liu, S. Suh, J. F. Vargas, B. Zhou, A. Grünerbl, and P. Lukowicz, "A wearable multi-modal edge-computing system for real-time kitchen activity recognition," in *International Joint Conference on Artificial Intelligence*, pp. 132–145, Springer, 2024.
- [37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [38] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.
- [39] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.
- [40] M. Ciliberto, V. F. Rey, A. Calatroni, P. Lukowicz, and D. Roggen, "Opportunity++: A multimodal dataset for video-, wearable, object and ambient sensors-based human activity recognition," 2021.
- [41] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th international symposium on wearable computers*, pp. 108–109, IEEE, 2012.
- [42] B. Barshan and M. C. Yükses, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *The Computer Journal*, vol. 57, no. 11, pp. 1649–1667, 2014.
- [43] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5137–5146, 2018.
- [44] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *bmvc*, vol. 2, p. 5, Aberystwyth, UK, 2010.
- [45] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.
- [46] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.
- [47] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [48] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, *et al.*, "Ai challenger: A large-scale dataset for going deeper in image understanding," *arXiv preprint arXiv:1711.06475*, 2017.

- [49] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11656–11665, 2021.
- [50] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.
- [51] S. Mukherjee, L. Anvitha, and T. M. Lahari, "Human activity recognition in rgb-d videos by dynamic images," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 19787–19801, 2020.
- [52] A. Tehrani, M. Yadollahzadeh-Tabari, A. Zehtab-Salmasi, and R. Enayatifar, "Wearable sensor-based human activity recognition system employing bi-lstm algorithm," *The Computer Journal*, vol. 67, no. 3, pp. 961–975, 2024.
- [53] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "Innohar: A deep neural network for complex human activity recognition," *Ieee Access*, vol. 7, pp. 9893–9902, 2019.
- [54] F. Xu, X. Gao, and W. Wang, "A human activity recognition model based on deep neural network integrating attention mechanism," *Scientific Reports*, vol. 15, no. 1, p. 23192, 2025.
- [55] A. Tsanousa, G. Meditskos, S. Vrochidis, and I. Kompatsiaris, "Multi-sensors for human activity recognition," 2023.
- [56] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10146–10176, 2014.
- [57] M. Ronald, A. Poulose, and D. S. Han, "isplinception: an inception-resnet deep learning architecture for human activity recognition," *IEEE Access*, vol. 9, pp. 68985–69001, 2021.
- [58] F. Wang, Y. Song, J. Zhang, J. Han, and D. Huang, "Temporal unet: Sample level human action recognition using wifi," *arXiv preprint arXiv:1904.11953*, 2019.
- [59] R. M. Narayanan and M. Zenaldin, "Radar micro-doppler signatures of various human activities," *IET Radar, Sonar & Navigation*, vol. 9, no. 9, pp. 1205–1215, 2015.
- [60] D. Roggen, A. Calatroni, L.-V. Nguyen-Dinh, R. Chavarriaga, and H. Sagha, "Opportunity activity recognition," 2010.
- [61] T. Pfizinger, M. Koch, F. Schlenke, and H. Wöhrle, "Multi-modal dataset of human activities of daily living with ambient audio, vibration, and environmental data," *Data*, vol. 9, no. 12, p. 144, 2024.

- [62] K. Liu, C. Gao, B. Li, and W. Liu, "Human activity recognition through deep learning: Leveraging unique and common feature fusion in wearable multi-sensor systems," *Applied Soft Computing*, vol. 151, p. 111146, 2024.
- [63] D. Madhuranga, R. Madushan, C. Siriwardane, and K. Gunasekera, "Real-time multimodal adl recognition using convolution neural networks," *The Visual Computer*, vol. 37, no. 6, pp. 1263–1276, 2021.
- [64] J. Roche, V. De-Silva, J. Hook, M. Moencks, and A. Kondoz, "A multimodal data processing system for lidar-based human activity recognition," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10027–10040, 2021.
- [65] A. A. Aguilera, R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo, "Multi-sensor fusion for activity recognition—a survey," *Sensors*, vol. 19, no. 17, p. 3808, 2019.
- [66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [67] A. Doering, D. Chen, S. Zhang, B. Schiele, and J. Gall, "Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20963–20972, 2022.
- [68] M. Verma, S. Kumawat, Y. Nakashima, and S. Raman, "Yoga-82: A new dataset for fine-grained classification of human poses," 2020.
- [69] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
- [70] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th international conference on mobile computing, applications and services*, pp. 197–205, IEEE, 2014.
- [71] M. Belal, T. Hassan, A. Ahmed, A. Aljarah, N. Alsheikh, and I. Hussain, "Feature fusion for human activity recognition using parameter-optimized multi-stage graph convolutional network and transformer models," in *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–7, IEEE, 2024.
- [72] D. Oved, I. Alvarado, and A. Gallo, "Real-time human pose estimation in the browser with tensorflow.js." Online Resource, TensorFlow Blog, 2018. <https://blog.tensorflow.org/2018/05/posenet-real-time-human-pose-estimation-in-browser.html>.
- [73] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

- [74] A. Sharma, Y. Shah, Y. Agrawal, and P. Jain, "Real-time recognition of yoga poses using computer vision for smart health care," *arXiv preprint arXiv:2201.07594*, 2022.
- [75] N. Giulietti, A. Caputo, P. Chiariotti, and P. Castellini, "Swimmernet: Underwater 2d swimmer pose estimation exploiting fully convolutional neural networks," *Sensors*, vol. 23, no. 4, p. 2364, 2023.
- [76] S. Dehdashtian, R. He, Y. Li, G. Balakrishnan, N. Vasconcelos, V. Ordonez, and V. N. Boddeti, "Fairness and bias mitigation in computer vision: A survey," *arXiv preprint arXiv:2408.02464*, 2024.
- [77] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
- [78] Y.-S. Chang, Y.-X. Zheng, and Z.-Y. Ku, "Comparative analysis of lightweight openpose and movenet ai models for real-time fall detection and alert systems," *Sensors & Materials*, vol. 37, 2025.
- [79] B. Jo and S. Kim, "Comparative analysis of openpose, posenet, and movenet models for pose estimation in mobile devices," *Traitement du Signal*, vol. 39, no. 1, p. 119, 2022.
- [80] G. Bradski, "The opencv library,," *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.
- [81] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous systems," 2015.
- [82] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *nature*, vol. 585, no. 7825, pp. 357–362, 2020.
- [83] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.
- [84] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [85] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

DECLARATION

I hereby certify that the work which is presented in the Major Project-II/Research Work entitled "**Multimodal Human Activity Recognition Using Deep Learning**" 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 my 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 me 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 Skeleton-Based Pose Estimation Models for Human Activity Recognition using CNN-LSTM

Author Names (in sequence as per research paper): Mansi Gupta, Dr. Abhilasha Sharma

Name of Conference/Journal: 7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (MIND-2025)

Conference Dates with venue (if applicable): 12-14 December 2025, Malaviya National Institute of Technology (MNIT), Jaipur, India

Have you registered for the conference?: Yes

Status of paper (Accepted/Published/Communicated): Accepted and Presented (Scopus-Indexed)

Date of paper Communication: 30 October 2025

Date of paper Acceptance: 18 November 2025

Date of paper Publication: To be published

24/SWE/09

MANSI GUPTA

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

Dr. Abhilasha Sharma

Date:

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).

Gupta, M. and Sharma, A., "*Comparative Analysis of Skeleton-Based Pose Estimation Models for Human Activity Recognition Using CNN-LSTM*". The paper has been **Accepted** as well as **Presented** at the **7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (MIND 2025)**, held at **MNIT Jaipur, India**, during **12-14 December 2025**. The conference proceedings are indexed by **Scopus**. Paper ID: **382**.

Submission:

7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences : Submission (382) has been created.

1 message

Microsoft CMT <noreply@msr-cmt.org>
To: mansigupta02ktt@gmail.com

Thu, Oct 30, 2025 at 11:36 PM

Hello,

The following submission has been created.

Track Name: AI and Intelligent Systems

Paper ID: 382

Paper Title: Comparative Analysis of Skeleton-Based Pose Estimation Models for Human Activity Recognition Using CNN-LSTM

Acceptance:

MIND 2025: ACCEPTANCE NOTIFICATION 382

2 messages

Microsoft CMT <noreply@msr-cmt.org>
To: Mansi Gupta <mansigupta02ktt@gmail.com>
Cc: mind2025.mnit@gmail.com

Tue, Nov 18, 2025 at 12:25 AM

Dear Mansi Gupta

Congratulations!!

We are pleased to inform you that your paper has been accepted for oral presentation (with revision) as a full paper for the 7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (MIND 2025) to be held at MNIT Jaipur, India, on 12-14 December 2025 in Hybrid mode.

Paper ID: 382

Paper Title: Comparative Analysis of Skeleton-Based Pose Estimation Models for Human Activity Recognition Using CNN-LSTM

Payment Proof:



7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences

MNIT Jaipur

Jaipur, Rajasthan - 302017

Email: mind2025@mnit.ac.in

Bill To

Mansi Gupta

Paper Id: 382



TAX INVOICE / RECEIPT

Invoice Number: MNIT/MIND2025-85

Date: 12/12/2025

Description	Amount(₹)
MIND 2025 Registration Fee:	
(Delegate/Presenter Category)	
GST @ 18%	
TOTAL	Rs 10620 (Incl. of all taxes)

Full payment received for MIND 2025, 12-14 December 2025, MNIT Jaipur.

Authorised Signatory: _____
Conference Chair
Finance Chair, MIND 2025 MIND 2025

Presentation Certificate:



Scopus Proof:

7th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences

MIND-2025

12-14 December 2025 | MNIT Jaipur, India

PAPER SUBMISSION <https://cmt3.research.microsoft.com/MINDS2025>

Conference Proceedings will be published by **Springer** in the **CCIS** series which is Scopus indexed.

Publishing Partner

Communications in Computer and Information Science, Springer (confirmed)

Latest Updates

- The attendee registration portal will be open soon.
- Registration deadline is now 22/11/25
- Final paper submission deadline is 30th October 2025. (CLOSED)
- The conference will be organized in **HYBRID** mode.
- MIND 2025 technically co-sponsored by IEEE - C I society Student chapter MNIT
- Paper submission has started.
- Proceedings publication in CCIS Springer is confirmed.

DECLARATION

I hereby certify that the work which is presented in the Major Project-II/Research Work entitled "**Multimodal Human Activity Recognition Using Deep Learning**" 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 my 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 me 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: A Comparative Analysis of Deep Learning Approaches for Human Pose Estimation: Benchmarks, Applications, and Future Outlook

Author Names (in sequence as per research paper): Mansi Gupta, Dr. Abhilasha Sharma

Name of Conference/Journal: 9th International Conference on Innovative Computing and Communication (ICICC-2026)

Conference Dates with venue (if applicable): 06-07 February 2026, Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, India

Have you registered for the conference?: Yes

Status of paper (Accepted/Published/Communicated): Accepted and Presented (Springer LNNS Proceedings, Scopus-Indexed)

Date of paper Communication: 15 November 2025

Date of paper Acceptance: 04 December 2025

Date of paper Publication: To be published

24/SWE/09

MANSI GUPTA

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

Dr. Abhilasha Sharma

Date:

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).

Gupta, M. and Sharma, A., "*A Comparative Analysis of Deep Learning Approaches for Human Pose Estimation: Benchmarks, Applications, and Future Outlook*". The paper has been **Accepted** as well as **Presented** at the **9th International Conference on Innovative Computing and Communication (ICICC-2026)**, held at **Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, India**, during **06-07 February 2026**. The conference proceedings are published in **Springer LNNS Series** and indexed by **Scopus**. Paper ID: **1052**.

Submission:

9th INTERNATIONAL CONFERENCE ON INNOVATIVE COMPUTING AND COMMUNICATION (ICICC-2026) : Submission (1052) has been created.
1 message

Microsoft CMT <noreply@msr-cmt.org> Sat, Nov 15, 2025 at 9:01 PM
To: mansigupta02kti@gmail.com

Hello,

The following submission has been created.

Track Name: SS_17: Applied Image Processing: Current Applications and Emerging Challenges

Paper ID: 1052

Paper Title: A Comparative Analysis of Deep Learning Approaches for Human Pose Estimation: Benchmarks, Applications, and Future Outlook

Acceptance:

ICICC 2026: Paper Notification for Paper ID 1052
2 messages

ICICC 2026 <icicc.ui@gmail.com> Thu, Dec 4, 2025 at 8:41 AM
To: Mansigupta02kti <mansigupta02kti@gmail.com>

9th International Conference on Innovative Computing and Communication (ICICC-2026) – A Flagship Conference

Dear Author(s),

Greetings from - ICICC 2026!

We congratulate you that your paper with submission ID '1052' and Paper Title '**A Comparative Analysis of Deep Learning Approaches for Human Pose Estimation: Benchmarks, Applications, and Future Outlook**' has been accepted for publication in the **Proceedings of ICICC 2026- Springer LNNS Series** [Indexing: zbMATH, Scopus, DBLP, EI Compendex & SCImago - Approved]. This acceptance means your paper is among the top 15% of the papers received/reviewed.

Payment Proof:



Presentation Certificate:



Scopus Proof:



About Conference

International Conference on Innovative Computing and Communication (ICICC-2026) is organized to bring together innovative scientists, professors, research scholars, students, and industrial experts in the field of Computing and Communication to a common forum. The primary goal of the conference is to promote the exchange of innovative scientific information between researchers, developers, engineers, students, and practitioners. Another goal is to promote the transformation of fundamental research into institutional and industrialized research and to convert applied innovation into the real-time application. Overall, the conference will provide the researchers and attendees with prospects for national and international collaboration and networking among universities and institutions from India and abroad for promoting research. ICICC-2026 will be held at Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi. All the accepted papers (after double-blind peer review) will be published as proceedings of ICICC-2026 in Springer LNCS Series (Approved).

[Paper Submission](#)

Downloads

DECLARATION

I hereby certify that the work which is presented in the Major Project-II/Research Work entitled “**Multimodal Human Activity Recognition Using Deep Learning**” 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 my 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 me for the award of any other degree of this or any other Institute/University. The work has been accepted for publication as a book chapter with the following details:

Title of the Chapter: Innovative Applications of Visual Perception in Modern AI Systems

Author Names (in sequence as per chapter): Dr. Roshni Singh, Mansi Gupta, Dr. Abhilasha Sharma

Name of Book: Advanced Techniques in Computer Vision: Current Research and Emerging Applications

Publisher: Bentham Science Publishers

Chapter Number: 14

Status of Chapter: Accepted for Publication

Date of Chapter Communication: 10 September 2025

Date of Chapter Acceptance: 06 May 2026

Date of Chapter Publication: To be published

24/SWE/09

MANSI GUPTA

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

Dr. Abhilasha Sharma

Date:

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).

Singh, R., Gupta, M. and Sharma, A., *"Innovative Applications of Visual Perception in Modern AI Systems"*. The chapter has been **Accepted for Publication** in the book titled **Advanced Techniques in Computer Vision: Current Research and Emerging Applications**, to be published by **Bentham Science Publishers**. Chapter Number: **14**.

Acceptance:

Acceptance of Chapter Submission – Innovative Applications of Visual Perception in Modern AI Systems

submission <callforarticles.submission@gmail.com>
to Roshni, me

May 6, 2026, 11:37 AM (1 day ago) ☆ ☺ ↶ ⋮

Dear Authors,

We are pleased to inform you that your chapter titled "Innovative Applications of Visual Perception in Modern AI Systems" submitted for the book "Advanced Techniques in Computer Vision: Current Research and Emerging Applications" has been accepted for publication.

Chapter Details: Chapter 14

Title: Innovative Applications of Visual Perception in Modern AI Systems

Authors: Roshni Singh, Mansi Gupta, and Abhilasha Sharma

Your chapter is currently under the publication process and will soon be submitted to Bentham Science Publishers for further processing and final publication.

We sincerely appreciate your valuable contribution to this book project and thank you for your cooperation throughout the submission process.

Scopus Proof:

CALL FOR BOOK CHAPTER

ADVANCED TECHNIQUES IN COMPUTER VISION: CURRENT RESEARCH AND EMERGING APPLICATIONS

**BENTHAM
SCIENCE**

TENTATIVE CHAPTERS

1. Introduction to Advanced Computer Vision
2. Machine Learning Foundations for Computer Vision
3. Deep Learning: The Backbone of Modern Computer Vision
4. Advanced Object Detection and Segmentation
5. 3D Vision and Spatial Understanding
6. Real-Time Vision Systems
7. Vision for Robotics and Automation
8. Computer Vision in Healthcare and Bioinformatics
9. Vision-Based Security and Surveillance
10. AI for Creativity: Generating Art, Music, and Beyond
11. Edge AI and IoT for Vision Systems
12. Emerging and Niche Applications in Computer Vision
13. Ethical and Societal Impacts of Computer Vision
14. AI-Powered Vision for Next-Gen Agriculture
15. Case Studies in Computer Vision
16. The Future of Vision AI

Indexed by
Scopus

**No Publication
Fee**

Contact Person:
Dr. Preksha Pareek

Phone: +91-7984611024

Email:
callforarticles.submission@gmail.com

DECLARATION

I hereby certify that the work which is presented in the Major Project-II/Research Work entitled “**Multimodal Human Activity Recognition Using Deep Learning**” 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 my 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 me 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: LiteHAR-DSC: A Depthwise Separable Convolutional Framework for Edge-Deployable Cross-User Human Activity Recognition on Wearable Sensors

Author Names (in sequence as per research paper): Mansi Gupta, Dr. Abhilasha Sharma, Dr. Roshni Singh

Name of Journal: Advanced Engineering Sciences

Conference Dates with venue (if applicable): NA

Have you registered for the conference?: NA

Status of paper (Accepted/Published/Communicated): Accepted

Date of paper Communication: 18 March 2026

Date of paper Acceptance: 29 May 2026

Date of paper Publication: Awaited

24/SWE/09

MANSI GUPTA

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

Dr. Abhilasha Sharma

Date:

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).

Gupta, M., Sharma, A., and Singh, R., "*LiteHAR-DSC: A Depthwise Separable Convolutional Framework for Edge-Deployable Cross-User Human Activity Recognition on Wearable Sensors*". The paper has been **Accepted** in the **Advanced Engineering Sciences** (ISSN: 2096-3246), published by **Sichuan University**. Article ID: **2026-643**.

Acceptance:

ARTICLE ACCEPTANCE LETTER

1 message

admin@advancedengineeringscience.com <admin@advancedengineeringscience.com>
To: mansi.gupta2026@gmail.com, abhisha.sharma@sdu.ac.in, roshni.singh_2621@du.ac.in

Fri, May 29, 2026 at 2:29 PM

May 29, 2026

ARTICLE ACCEPTANCE LETTER

Article ID: 2026-643

Article Title: LiteHAR-DSC: A Depthwise Separable Convolutional Framework for Edge-Deployable Cross-User Human Activity Recognition on Wearable Sensors

Authors: Mansi Gupta, Abhisha Sharma, Roshni Singh

Thank you very much for your submission to our journal.

We are pleased to inform you that your paper has been reviewed and accepted for publication. Your article will be published in upcoming current issue.

Thank you for making the journal a vehicle for your research interests.

Best wishes,

Editor-in-Chief

Advanced Engineering Sciences

ISSN: 2096-3246

<https://advancedengineeringscience.com/>
<https://www.scopus.com/sourceid/21100805730>
<https://www.scimagr.com/journalsearch.php?q=21100805730&tp=sk&dian=0>

Scopus Proof:

[scopus.com/sourceid/21100805730](#)

Scopus Preview

[Author Search](#)
[Sources](#)
[Create account](#)
[Sign in](#)

Source details

[Feedback](#)
[Compare sources](#)

Gongcheng Kexue Yu Jishu/Advanced Engineering Sciences

Formerly known as: [Sichuan Daxue Xuebao \[Gongcheng Kexue Ban\]](#)/Journal of Sichuan University (Engineering Science Edition)

Years currently covered by Scopus: from 2017 to 2026

Publisher: [Sichuan University](#)

ISSN: [2096-3246](#)

Subject area: [Engineering: General Engineering](#)

Source type: [Journal](#)

[View all documents](#)
[Set document alert](#)
[Save to source list](#)

CiteScore 2024	2.2
SJR 2024	0.304
SNIP 2024	0.665

[CiteScore](#)
[CiteScore rank & trend](#)
[Scopus content coverage](#)

CiteScore 2024

2.2

1,357 Citations 2021 - 2024
 621 Documents 2021 - 2024

Calculated on 05 May, 2025

CiteScoreTracker 2025

2.5

1,627 Citations to date
 661 Documents to date

Last updated on 05 April, 2026 • Updated monthly



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, 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)

AUTHOR'S PROFILE

Mansi Gupta is currently pursuing the Master of Technology (M.Tech.) degree in Software Engineering from the Department of Software Engineering, Delhi Technological University, Delhi, India. Her research work focuses on deep learning approaches for Human Activity Recognition (HAR), Human Pose Estimation (HPE), multimodal sensor fusion, and edge-oriented intelligent systems. Her thesis entitled "*Multimodal Human Activity Recognition Using Deep Learning*" investigates pose estimation architectures, skeleton-based activity recognition, and multimodal temporal learning frameworks for robust cross-user activity recognition.



Her research interests include Computer Vision, Deep Learning, Human Activity Recognition, Human Pose Estimation, Multimodal Learning, Wearable Sensor Analytics, Transformer-based Architectures, and Edge Artificial Intelligence. During her post-graduate research, she has worked on skeleton-based HAR using CNN-LSTM fusion architectures and multimodal transformer frameworks integrating wearable and ambient sensor streams.

She has contributed to research publications including international conferences, journal submissions, and book chapters in the domains of human activity recognition, pose estimation, and computer vision. Her research contributions include comparative analysis of deep learning-based pose estimation approaches, evaluation of multiple pose estimation backends for skeleton-based HAR, and the proposed MM-HTT framework for cross-user multimodal activity recognition.