Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Reproducible and Generalizable Multimodal HAR Using Attention-Based Fusion and Domain Adaptation

Swati Gautam , Ankush Srivastava
Department of CSE, Ram Krishna Dharmarth Foundation University, Gandhi Nagar, Bhopal (M.P.)

Abstract*: *This work presents a reproducible and multimodal framework for human activity recognition (HAR) that addresses challenges related to domain adaptation, domain shift, and inconsistent evaluation protocols. The proposed approach combines modality- specific encoders, cross-modal co-attention (CMCA), and domain adversarial neural networks (DANN) to learn adaptive, domain-invariant representations. Furthermore, a reproducibility module is included to ensure uniform preprocessing and evaluation using a strict leave-one-subject-Out (LOSO) protocol. The performance of the proposed model is evaluated on the UTD-MHAD, PAMAP2, and Opportunity datasets. Experimental results show consistent improvement over baseline approaches, with a maximum accuracy gain of 6.2% on the Opportunity dataset compared to early fusion. The findings further indicate reduced inter-class confusion and improved robustness to noise and degradation of the modalities.

Keywords*: *Human Activity Recognition (HAR), Multimodal Learning, Cross-Modal Co-Attention, Domain Adaptation, Reproducibility, Generalization.

I. INTRODUCTION

Human activity recognition (HAR) plays a vital role in intelligent systems for various applications such as healthcare, human-computer interaction, and behavioral analysis, where there is a need for automated recognition of human actions based on sensor data [1][2]. The majority of existing HAR methodologies employ single-mode input modalities like wearables IMUs or video RGB stream [3]. Such techniques may suffer from low robustness in realistic scenarios because of sensor noise, occlusion, changing lighting conditions, and subject-to-subject variation [4].

The recent trend in HAR has been towards multimodal approaches where not only appearance but also motion-related information extracted from various sensors is incorporated into the system[5]. RGB videos, skeleton data, and inertial measurement units (IMU) data have been considered for constructing multimodal HAR [6]. Moreover, deep neural networks such as CNNs, RNNs, and Transformers have gained importance in the field. In this regard, the emerging paradigm of reproducible and generalizable multimodal HAR is intended to ensure consistent behavior in different users and deployment environments[7].

Despite all the progress mentioned above, there are persistent problems in multimodal HAR algorithms, namely reproducibility, cross-modal consistency, and generalization to unseen subjects and domains. These problems have been compounded further due to non-standardization in the pre-processing stage, inconsistent evaluation approaches, and poor handling of modality failure. As a result, the design of unified, reproducible, and generalizable frameworks to support reliable multimodal learning and cross-benchmark evaluation is a current research task. In this paper, these problems are addressed by introducing a multimodal HAR framework, which consists of a standardized preprocessing pipeline, adaptive CMCA fusion, and domain-adversarial representation learning.

II. RELATED WORK

The design of accurate HAR models has been increasingly dependent on the use of heterogeneous sensor fusion and neural network architecture for handling real-life scenarios. In [8], a self-supervised pre-training scheme utilizing masking of synchronized data was proposed to extract spatiotemporal information without relying on external annotations. In [9], a model was proposed for optimizing action recognition by integrating residual convolutions with attention mechanisms based on hierarchy and grammatical matrices. To mitigate the problem of poor data quality, [10] proposed a denoising autoencoder with self-attention for handling multimodal correlation across sensors. To resolve the problem of high computational requirements associated with multimodal models in edge computing environments, [11] designed an approach utilizing knowledge distillation for transferring features from teacher to student models.

A multistream factorized transformer model was developed in [12] by leveraging temporal embeddings to exploit local variations and alignment without feature extraction modules. Foundation models are also another approach to accommodate different modalities, which includes a model that uses masked data modeling and few-shot alignment with respect to sensor input variations and devices proposed in [7]. Deep reinforcement learning and large language models' contributions to HAR were reviewed in detail in [13]; while in [14], a technique using confidence-based gradient adjustment was suggested as a solution to mitigate modality conflict problems. The development of sensor-based and hybrid methods was investigated in [15].

Generalization among various users continues to be an obstacle owing to subject-specific biomechanics. The authors

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

in [16], addressed the challenge by suggesting the use of a Siamese adversarial network with the goal of minimizing variations between representation models of various subjects. To ensure generalization throughout the training phase, [17] offers a review of tunable parameters, suggesting standardized approaches to HAR studies. The authors of [18], introduces a graph-based model augmented by edge information, which is able to leverage domain invariant characteristics by using anatomical correlation information. In [19], it is proven that including discriminative activity-based subjects in adversarial networks improves classification performance considerably.

The concepts of efficiency and alignment based on attention mechanisms have been further improved in recent literature to allow reproducibility. In [20], a lightweight residual network architecture for inertial signals was designed to provide optimal performance versus complexity trade-off. The work in [21], addressed the problem of inter-subject variability for embedded solutions by demonstrating that embedding the variability in modeling activities performed by different subjects increases the robustness of models. The authors in [22], introduced a unified contrastive fusion transformer and a factorized time-modality transformer to achieve efficient precision fusion in the sense of complexity and contextual representation richness. Lastly, the authors in [23], developed an approach using inter-segment attention mechanism and converting the inertial signal into Gramian angular fields to capture activity properties efficiently.

III. PROBLEM FORMULATION

HAR involves inferring human actions from multimodal time-series data collected by different sensors. Here, the problem of HAR is formulated as that of multi-modal sequence learning under domain shift and noise perturbation.

A. Temporal Alignment and Multimodal Representation

Let $\mathcal{M} = \{M_1, M_2, \dots, M_k\}$ denote a set of K heterogeneous sensing modalities such as IMUs, RGB video, and skeletal pose. Each modality M_k produces a time-series signal:

$$X_k = \{x_k^t\}_{t=1}^{T_k}, \quad X_k \in \mathbb{R}^{T_k \times D_k} \tag{1}$$

where T_k and D_k denote the temporal length and feature dimensionality, respectively. Considering the different sampling rates of the raw data, they are first synchronized in terms of temporal resolution to T using synchronization function $S(\cdot)$:

$$\hat{X}_k = S(X_k), \hat{X}_k \in \mathbb{R}^{T \times D_k} \tag{2}$$

The synchronized modality is then transformed to latent representation as follows:

$$Z_k = \phi_k(\hat{X}_k), \qquad Z_k \in \mathbb{R}^{T \times d} \tag{3}$$

where $\phi_k(\cdot)$ maps modality-specific spatial-temporal information into unified space having dimensionality d. The learned fusion function Ψ is used for aggregation to get fused features Z as below:

$$Z = \Psi(Z_1, Z_2, \dots, Z_k) \tag{4}$$

B. Learning Objective and Domain Invariance

The model is trained to minimize an objective which considers both activity recognition and domain invariance together. The total loss L is defined as:

$$L = L_{cls}(\hat{y}, y) + \lambda L_{domain}(Z) \tag{5}$$

where:

- L_{cls} is the classification loss based on supervised training data.
- L_{domain} is the loss incurred by domain-adversarial network calculated through Gradient Reversal Layer. It helps to achieve invariance of subjects' features but still preserves activity discriminative features.
- λ is a hyperparameter which determines the balance between classification and domain invariance.

C. Robustness to Noise and Perturbations

In practical situations, the sensor readings might be subjected to noise contamination or failure. The contaminated input data can be described in the form $\tilde{X}_k = \hat{X}_k + \varepsilon_k$, where ε_k denotes the noise in terms of stochastic nature or missing data regions. The system is expected to comply with the robustness condition:

$$\| F(X) - F(\tilde{X}) \| < \delta \tag{6}$$

where F is the mapping function and δ is a predetermined tolerance level.

D. Generalization under Domain Shift

To achieve generalization to unseen users, a Leave-One-Subject-Out (LOSO) framework is adopted. Considering a source domain P_s and a target domain P_t such that $P_s \neq P_t$, the aim is such that the learned features Z maintain their invariance with respect to the subject identity but at the same time be discriminative about the target tasks.

E. Reproducibility Constraint

To resolve the reproducibility crisis in HAR, the suggested approach complies with the consistency principle. Here, experimental results R can be represented as a function of the predefined pipeline $P(\cdot)$, the model structure A, and the evaluation function $E(\cdot)$.

$$R = E(F(P(X); A)) \tag{7}$$

The above-mentioned parameters remain constant, and open-source preprocessing scripts are used to prove the reproducibility of the results.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

IV. PROPOSED METHODOLOGY

This research proposes a multimodal HAR framework designed to improve robustness, generalization, and experimental consistency. It addresses three problems of domain shift, sensor noise, and the absence of a common evaluation protocol by incorporating noise aware feature extraction, cross modal co attention, and adversarial domain adaptation into one framework.

A. Multimodal Input Representation

The input vector is denoted by $X = \{X_1, X_2, \ldots, X_k\}$, where X_k refers to the input vectors of time series collected using different sensors, such as IMUs, RGB videos, and skeletal poses. To maintain the consistency in terms of time for each stream despite the different sampling rates used, the following normalization function is employed:

$$\hat{X}_k = S(X_k) \tag{8}$$

The function performs linear interpolation to match all modalities in terms of their temporal granularity in order to ensure synchronization before further computations. Considering the differences in the availability of different modalities from one data set to another, this model is designed based on the use of only those modalities that are available, without generating any additional input

B. Noise-Aware Preprocessing via Denoising Autoencoder

In order to simulate the effects of sensor noise, as well as possible signal dropout during training time, the proposed framework uses a preprocessing phase for handling sensor noise. Stochastic noise is added to each modality [24]:

$$\tilde{X}_k = \hat{X}_k + \epsilon_k, \qquad \epsilon_k \sim N(0, \sigma^2) \tag{9}$$

The denoising autoencoder (DAE) is then used to denoise and restore the original signal:

$$X_k^{clean} = DAE(\tilde{X}_k) \tag{10}$$

The optimization objective of the framework is defined as follows:

$$L_{rec} = \sum_{k=1}^{K} \| \hat{X}^k - X_k^{clean} \|^2 \tag{11}$$

It ensures that the proposed framework learns noise-invariant signatures before entering the feature encoding step.

C. Modality-Specific Feature Encoding

The processed inputs are passed through modality-specific encoders $\phi_k(X_k^{clean})$. In case of IMU signals, conversion into gramian angular field (GAF) images is applied to maintain the temporal-spatial correlations within the signals and processing is done through CNNs. Visual streams pass through transformers to detect long-range spatial-temporal dependencies. To tackle the issue of dimensional mismatch between different sensors, the following transformation is used for mapping [25]:

$$\tilde{Z}_k = W_k Z_k + b_k \tag{12}$$

where W_k is a learned projection matrix that projects varying dimensions into a common d-dimension.

D. Adaptive CMCA Fusion

The proposed CMCA mechanism models interactions among available modalities. In case only one modality is present, the attention mechanism gives all its weight to the available modality, while the fusion process becomes an identity function. For each modality:

$$Q_k = W_q \tilde{Z}_k, \quad K_k = W_k^{att} \tilde{Z}_k \tag{13}$$

The attention score is calculated via the interaction as:

$$\gamma_k = \frac{Q_k^{\top} K_k}{\sqrt{d}}, \qquad \alpha_k = softmax(\gamma_k) \tag{14}$$

The fused representation $Z_{fused} = \sum_{k=1}^{K} \alpha_k \tilde{Z}_k$, allows the model to prioritize reliable sensors.

E. Domain-Invariant Feature Learning (DANN)

To mitigate performance degradation introduced by the domain shift problem, a DANN is incorporated. The goal of the domain classifier $D(\cdot)$ is to predict the domain ID of the subject using the fusion features, and the GRL encourages the feature extractor to learn subject-invariant features:

$$L_{adv} = -\mathbb{E}_{x_s}[log(D(Z_{fused}^s))] - \mathbb{E}_{x_t}[log(1 - D(Z_{fused}^t))] \tag{15}$$

The activity classifier G_y performs classification of the activities with a cross-entropy loss, L_{cls}.

F. Joint Optimization Objective

The objective function defines a multitask learning formulation:

$$\min_{\phi_k, G_y} \max_{D} L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{rec} \tag{16}$$

where λ_1 regulates the extent of the doman-invariance and λ_2 ensures the robustness of the signal reconstruction.

The overall architecture of the proposed framework, which involves multimodal feature encoding, cross-modal adaptive attention fusion, domain adversarial learning, and the repeatability component, is shown in Figure 1.

V. COMPUTATIONAL COMPLEXITY

Computational complexity results from feature representation, co-attention fusion, and domain adversarial learning Let K be the number of input modalities, and d be the size of the feature vector. The computational complexity of the co-attention model is $O(K \cdot d^2)$ which is analogous to the conventional models utilizing the attention mechanism. Domain adversarial learning is relevant only to the training stage and does not impose much complexity on inference. The overall computational complexity of the framework is moderate compared to existing multimodal methods.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

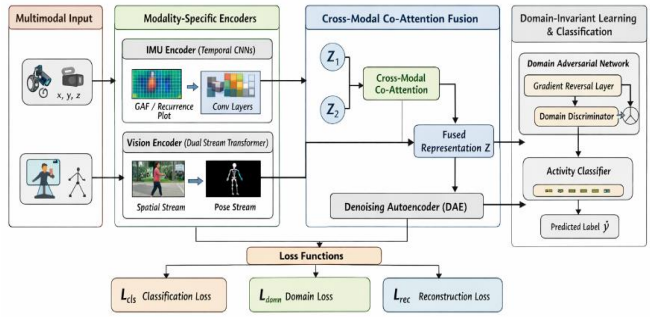

Figure 1. Overview of the proposed multimodal HAR framework.

VI. RESULTS AND DISCUSSION

In this section, experimental setup and evaluation metrics are described first, followed by a detailed discussion of the results.

A. Experimental Setup and Metrics

To evaluate the effectiveness of the proposed approach, the experiments were carried out on three popular datasets: UTD-MHAD [26], PAMAP2 [27], and Opportunity [28]. The UTD-MHAD is used for multimodal evaluation, whereas PAMAP2 and Opportunity are evaluated under IMU or multi-sensor configurations depending on available modalities. In order to determine the generalization capability of the model, a LOSO cross validation strategy was adopted. Two performance metrics such as Accuracy and Macro-F1 score are used to evaluate the results, while the latter is especially important due to the class imbalance characteristic of the Opportunity data set. The performance evaluation of proposed approach is carried out under both multimodal and sensor dominant conditions in case of availability of the modalities for adaptability to different modalities. The implementation details are summarized in Table-I.

Table- I: Implementation Details

Item	Description
Framework	PyTorch 2.1
Hardware	NVIDIA RTX 3090 GPU
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
Batch Size	64
Epochs	80
Windowing	2.56 s (UTD-MHAD), 4 s (PAMAP2), 50% overlap
Data Augmentation	Gaussian noise ($\sigma^2 = 0.01$)
Regularization	Dropout (0.5), weight decay (1×10^{-5})
Loss Weights	$\lambda_1 = 0.5$, $\lambda_2 = 0.3$
Random Seed	42
Preprocessing	Normalization, synchronization, segmentation
Evaluation Metrics	Accuracy, Macro-F1

B. Quantitative Results

Table-II summarizes the comparative performance of the models on UTD-MHAD, PAMAP2, and Opportunity datasets. The LOSO protocol is used to evaluate cross-subject generalization in realistic scenarios where the system needs to adapt to unknown users.

Table- II: Performance Comparison across Benchmarks

Dataset	Method	Accuracy (%)	F1 score (%)
UTD-MHAD	Single Modality (IMU)	88.4	87.1
	Early Fusion (Concat)	91.2	90.5
	Proposed (Co-Attn + DANN)	94.8	94.2
PAMAP2	Single Modality (IMU)	85.2	84.6
	Early Fusion (Concat)	87.5	86.9
	Proposed (Co-Attn + DANN)	91.3	90.8
Opportunity	Single Modality (IMU)	72.1	68.4
	Early Fusion (Concat)	76.5	73.2
	Proposed (Co-Attn + DANN)	82.7	80.5

As demonstrated, the proposed method shows improved performance compared to both the single modality baseline and the early fusion approach on all datasets. On the UTD-MHAD dataset, the performance of the model is measured in terms of accuracy at 94.8%, which is a 3.6% increase compared to the early fusion approach. On the PAMAP2 dataset, a similar increment of 3.8% is also attained. The model is also capable of providing better performance on the Opportunity dataset, where an increase of 6.2% in accuracy from 76.5% to 82.7% is achieved.

C. Ablation Study: Impact of DANN and Co-Attention

To evaluate the contribution of individual modules, an ablation study was performed using the UTD-MHAD dataset. The result presented in Table-III shows that the contributions of co-attention and domain adaptation are complementary.

Table- III: Ablation Study on UTD-MHAD

Configuration	Accuracy (%)	F1-Score (%)	Difference (vs. Base)
Base (Concat)	91.2	90.5	-
Base + Co-Attention	92.7	92.1	+1.5%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Base + DANN	93.1	92.5	+1.9%
Full Framework (Co-Attn + DANN)	94.8	94.2	+3.6%

Based on the findings, it is observed that DANN offers a notable improvement for the LOSO setup, as it encourages the feature extractor to extract features that are robust to any style of walking/motion. At the same time, Co-Attention tackles heterogeneous data by assigning modality weights dynamically.

D. Qualitative Analysis: Fusion Interpretability

Qualitative analysis of attention weights (α_k) indicates that the proposed framework is efficient in handling sensor reliability issues on-the-fly. When faced with modality degradation scenarios, such as a subject hiding behind furniture, the Co-Attention layer was noted to assign higher attention weights (up to 0.85) to the IMU sensor channel. On the other hand, when IMU sensors experience drift, the attention shifts toward more reliable modalities.

Additionally, the addition of the Reproducibility Layer solves one of the major challenges in HAR studies by ensuring standardization of the evaluation pipeline. By adopting stringent LOSO splits and signal-to-image conversions GAF, the proposed framework supports consistent and reproducible evaluation of the findings, irrespective of body morphology differences.

VII. LIMITATIONS

Although there is a noted level of performance, there are still some limitations that need to be addressed. First, the experiments are only carried out using a selected few open-source datasets, namely UTD-MHAD, PAMAP2, and Opportunity. This may fail to capture the true extent of diversity found in real-world data. Second, not all the datasets offer a complete combination of modalities, making it necessary to carry out experiments in multimodal and sensor-heavy configurations. Third, the co-attention module requires higher computation power than regular fusion modules.

VIII. CONCLUSION

In this paper, we propose a multimodal framework for human activity recognition based on modality-dependent representation, cross-modal co-attention mechanism, and domain-independent learning. The experimental results obtained on the UTD-MHAD, PAMAP2, and Opportunity datasets, demonstrate improved performance over the baseline techniques using the LOSO protocol. This implies that there is enhanced generalization capability, less confusion among classes, and more robustness to noise and degrading modality. Future research will concern further development of the approach using different datasets and optimizing its computational performance.

FUNDING INFORMATION

This work did not receive any specific funding from any funding agency, whether in the public, commercial, or not-for-profit sector.

ETHICS STATEMENT

This work did not involve human or animal participants and, therefore, did not require approval from an ethics committee.

STATEMENT OF CONFLICT OF INTERESTS

The authors of this work do not have any conflicts of interest to report.

REFERENCES

[1] T. Qureshi, M. Shahid, 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, 2025, doi: 10.1007/s10462-025-11275-x.

[2] H. Dong *et al.*, “Advances in Multimodal Adaptation and Generalization: From Traditional Approaches to Foundation Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, 2025, doi: 10.1109/TPAMI.2026.3651319.

[3] T.-H. Le, T.-K. Nguyen, T.-K. Tran, T.-H. Tran, and C. Pham, “GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer,” in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 297–303. doi: 10.1109/APSIPAASC58517.2023.10317315.

[4] A. S. M. Miah, Y. S. Hwang, and J. Shin, “Sensor-Based Human Activity Recognition Based on Multi-Stream Time-Varying Features With ECA-Net Dimensionality Reduction,” *IEEE Access*, vol. 12, pp. 151649–151668, 2024, doi: 10.1109/ACCESS.2024.3473828.

[5] Z. Quan *et al.*, “SMTDKD: A Semantic-Aware Multimodal Transformer Fusion Decoupled Knowledge Distillation Method for Action Recognition,” *IEEE Sensors Journal*, vol. 24, no. 2, pp. 2289–2304, 2023, doi: 10.1109/jsen.2023.3337367.

[6] P. Guo and M. Nakayama, “Towards User-Generalizable Wearable-Sensor-Based Human Activity Recognition: A Multi-Task Contrastive Learning Approach,” *Sensors*, vol. 25, no. 22, p. 6988, Nov. 2025, doi: 10.3390/s25226988.

[7] G. Zhu *et al.*, “MASTER: A Multi-modal Foundation Model for Human Activity Recognition,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 9, no. 3, Sep. 2025, doi: 10.1145/3749511.

[8] R. Liu and X. Liu, “MU-MAE: Multimodal Masked Autoencoders-Based One-Shot Learning,” in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2024, pp. 253–259. doi: 10.1109/MIPR62202.2024.00048.

[9] E. H. Houssein, I. A. Ibrahim, M. A. Mahdy, M. Kayed, A. M. Albarrak, and W. M. Mohamed, “Optimizing action recognition: a residual convolution with hierarchical and gram matrix based attention mechanisms,” *Journal Of Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01293-5.

[10] S. Xaviar, X. Yang, and O. Ardakanian, “Centaur: Robust Multimodal Fusion for Human Activity Recognition,” *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18578–18591, Jun. 2024, doi:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

10.1109/jsen.2024.3388893.

[11] J. Li, L. Yao, B. Li, and C. Sammut, "Distilled Mid-Fusion Transformer Networks for Multi-Modal Human Activity Recognition," *arXiv (Cornell University)*, 2023, doi: 10.48550/arxiv.2305.03810.

[12] X. Zhou, J. Yuan, L. Fan, X. Niu, K. Zha, and X. Liu, "MSMFT: Multi-Stream Multimodal Factorized Transformer for Human Activity Recognition," *IEEE Sensors Journal*, vol. 25, no. 6, pp. 10402–10416, 2025, doi: 10.1109/jsen.2025.3529889.

[13] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2018, doi: 10.1016/j.patrec.2018.02.010.

[14] P. Ji, J. Song, Y. Lu, H. Xiao, H. Liu, and C. Li, "Confidence-driven Gradient Modulation for Multimodal Human Activity Recognition: A Dynamic Contrastive Dual-Path Learning Approach," *arXiv (Cornell University)*, 2025, doi: 10.48550/arxiv.2507.02826.

[15] 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, Mar. 2025, doi: 10.3390/jimaging11030091.

[16] L. Bai, L. Yao, X. Wang, S. S. Kanhere, B. Guo, and Z. Yu, "Adversarial Multi-view Networks for Activity Recognition," *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–22, 2020, doi: 10.1145/3397323.

[17] Y. Huang, H. Zhao, Y. Zhou, T. Riedel, and M. Beigl, "Standardizing Your Training Process for Human Activity Recognition Models – A Comprehensive Review in the Tunable Factors," *Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 15–27, 2024, doi: 10.1007/978-3-031-63992-0_2.

[18] X. Ye and K. I. Wang, "Domain-Adversarial Anatomical Graph Networks for Cross-User Human Activity Recognition," *ArXiv.org*, 2025, doi: 10.48550/arxiv.2505.06301.

[19] F. M. Calatrava-Nicolás, S. Miyauchi, and O. M. Mozos, "Deep Adversarial Learning with Activity-Based User Discrimination Task for Human Activity Recognition," *arXiv (Cornell University)*, 2024, doi: 10.48550/arxiv.2410.12819.

[20] F. M. Calatrava-Nicolás and O. M. Mozos, "Light Residual Network for Human Activity Recognition using Wearable Sensor Data," *IEEE Sensors Letters*, vol. 7, no. 10, pp. 1–4, 2023, doi: 10.1109/LSENS.2023.3311623.

[21] F. M. Calatrava-Nicolás, S. Miyauchi, V. F. Rey, P. Lukowicz, T. Stoyanov, and O. M. Mozos, "Embedded Inter-Subject Variability in Adversarial Learning for Inertial Sensor-Based Human Activity Recognition," in *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2025, pp. 1–6. doi: 10.1109/MLSP62443.2025.11204225.

[22] K. O. Yang, J. Koh, and J. W. Choi, "UCFFormer: Recognizing human actions from multimodal sensors using unified contrastive fusion transformer," *Neurocomputing*, vol. 655, p. 131374, 2025, doi: 10.1016/j.neucom.2025.131374.

[23] Z. Gao *et al.*, "MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 3, Sep. 2023, doi: 10.1145/3610872.

[24] J. Yi and Z. Chen, "Variational Mixture of Stochastic Experts Auto-Encoder for Multi-Modal Recommendation," *IEEE Transactions on Multimedia*, vol. 26, pp. 8941–8954, 2024, doi: 10.1109/TMM.2024.3384058.

[25] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[26] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 168–172. doi: 10.1109/ICIP.2015.7350781.

[27] A. Reiss, "PAMAP2 Physical Activity Monitoring." UCI Machine Learning Repository, 2012. doi: 10.24432/C5NW2H.

[28] D. Roggen *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, 2010, pp. 233–240. doi: 10.1109/INSS.2010.5573462.

AUTHORS PROFILE

Mrs. Swati Gautam is pursuing a Ph.D. in Computer Science and Engineering at RKDF University, Bhopal, Madhya Pradesh, India. She completed her B. Tech. and M. Tech. in Computer Science and Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), Bhopal, Madhya Pradesh, India. Her areas of interest are human activity recognition, deep learning, neural networks, and machine learning.

Dr. Ankush Shrivastava completed his B.Tech. in Computer Science and Engineering and M.Tech. (Hons.) in Computer Science and Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), Bhopal, India. He completed his Ph.D. in Computer Science and Engineering from RKDF University, Bhopal, India. He is working as an Associate Professor in RKDF University, Bhopal. His research area includes wireless sensor networks, network security, and machine learning. His Ph.D. research work was on the recognition and prevention of multiple attacks in wireless sensor networks using neural networks and on-demand multicast routing protocols.