International Journal of Theoretical and Applied Computational Intelligence
https://ijtaci.com

Research Article

Human Action Recognition: A Comprehensive Survey of Multimodal Advances, Challenges, and Emerging Directions

Umar Suleiman Bichi[1] and Sunusi Bala Abdullahi[2]

[1] Department of Computer Science & Engineering, Faculty of Engineering and Technology, Vivekananda Global University, Jaipur, India
[2]Department of Information System Engineering, Faculty of Computer and Information Sciences, Sakarya University, Turkey
[*]Corresponding Author: Sunusi Bala Abdullahi (ORCID: https://orcid.org/0000-0003-1898-7352)
Received: 5/9/2025; Accepted: 20/10/2025; Published: 28/10/2025
https://doi.org/10.65278/IJTACI.2025.36

Abstract: Human Action Recognition (HAR) has emerged as a pivotal domain within computer vision and machine learning, driven by its transformative potential across surveillance, healthcare, human-computer interaction, and sports analytics. Despite notable advances, a persistent gap remains between benchmark-driven performance and real-world applicability, particularly in scenarios demanding cross-subject generalization, fine-grained understanding, computational, and scalability. This survey presents a systematic and critical review of HAR research published between 2022 and 2025, encompassing 30 peer-reviewed articles from the IEEE Xplore digital library. We trace the progression from unimodal frameworks to multimodal fusion architectures, highlighting innovations across skeleton-based, sensor-based, and vision-based modalities. Key architectural trends include transformer-based models, graph neural networks, and self-supervised learning, alongside domain-specific adaptations in healthcare and sports. Furthermore, we examine methodological shifts toward lightweight and generalizable systems. By synthesizing these developments, this work offers a structured roadmap for future research, emphasizing the need for robust evaluation protocols, ethical considerations, and deployment-ready HAR solutions.

Keywords: Human Action Recognition; Computer Vision, Cross-Subject Generalization, Deep Learning, Multimodal Fusion; Sensor-Based Activity Recognition; Privacy-Preserving Models; Transformer Architectures; Graph Neural Networks

1. Introduction

Human Activity Recognition (HAR) has emerged as a pivotal field within artificial intelligence (AI) and ubiquitous computing, driven by applications in healthcare, human–computer interaction, and smart environments. Over the past decade, researchers have increasingly modeled human activity using diverse modalities, ranging from vision-based systems to wearable sensors [1]. Despite substantial progress, significant challenges remain, including limited large-scale datasets [2], privacy-preserving learning [3], and poor generalization across demographic groups [4]. Deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph convolutional networks (GCNs) dominate HAR research due to their ability to capture temporal and spatial dependencies [5]. More recently, transformer-based methods have also demonstrated strong performance in multimodal HAR [6][25].

To contextualize current progress, this review provides a structured synthesis of recent HAR approaches, public datasets, evaluation metrics, and deployment challenges. In particular, we highlight gaps in dataset diversity, federated learning for privacy, and explainable AI for clinical and human–robot interaction applications. Unlike prior surveys, we integrate both technical developments and real-world considerations to provide a roadmap for future HAR research.

By systematically analyzing a curated collection of 30 research papers published between 2022 and 2025, this review focuses on three key areas: The evolution from single-modality systems to advanced multimodal fusion techniques [7]; The shift from general action recognition to specialized, fine-grained, and domain-specific applications [8][9]; and the emerging challenges of efficiency, privacy, and real-world generalization[10][5].

The remainder of this review is structured as follows: Section II, "Materials and Methods," describes the systematic approach adopted for identifying, selecting, and analyzing the reviewed studies, including the search strategy, inclusion criteria, and data extraction process. Section III, Summary of Key Observations, presents the main findings, covering HAR system architecture, application areas, datasets, algorithmic techniques, and major research challenges. Section IV highlights the open issues. Section V, Conclusion and Future Work, highlights the overall insights gained and outlines future directions for advancing research in Human Action Recognition (HAR).

2. Materials and Methods

This section outlines the methodological framework used to conduct a systematic review of Human Action Recognition (HAR) systems. The objective was to ensure comprehensive coverage and a rigorous evaluation of relevant research published between 2022 and 2025, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [1]. Searches were conducted primarily in IEEE Xplore, and supplemented with ACM Digital Library, Scopus, and Web of Science to avoid database bias. The selection process, depicted in Figure 1, adheres to PRISMA standards.

This review was conducted in accordance with PRISMA guidelines [1], encompassing studies published between 2022 and 2025. Searches were performed in IEEE Xplore, ACM Digital Library, Scopus, and Web of Science to ensure comprehensive coverage. The selection process is shown in Figure 1.

2.1 Search Strategy

A structured search strategy was implemented using keywords such as: "human action recognition", "activity recognition", "multimodal fusion", "pose estimation", "graph convolutional networks (GCN)", "transformers", "few-shot learning", and "cross-subject generalization". Boolean operators (AND/OR) were applied, and grey literature sources (e.g., arXiv) were screened.

2.2 Inclusion And Exclusion Criteria

The inclusion and exclusion criteria were precisely defined to select studies that significantly contributed to the current understanding of HAR systems and to ensure the academic rigor and relevance of this review.

2.2.1 Inclusion Criteria:

1. Peer-reviewed articles published in English between 2022 and 2025.
2. HAR studies using RGB, skeleton, Inertial Measurement Unit (IMU), or multimodal data.
3. Publications in IEEE venues
4. Studies with accessible full text that proposed novel architectures, datasets, or addressed key challenges like fine-grained recognition, cross-domain generalization, or computational efficiency.

2.2.2 Exclusion Criteria:

1. Studies published before 2022 or from sources other than IEEE Xplore.
2. Duplicate studies.
3. Review articles and meta-analyses, which were used for background context but not included in the final synthesis.
4. Non-peer-reviewed or inaccessible full texts

2.3 Selection Process

The initial search yielded 1,245 records. After removing 145 duplicates, 1,100 records remained. Screening excluded 800 based on titles/abstracts. Of the 300 full-text articles assessed, 225 were excluded for lack of methodological rigor. A total of 30 studies were included in the qualitative synthesis, and 45 in the quantitative analysis.

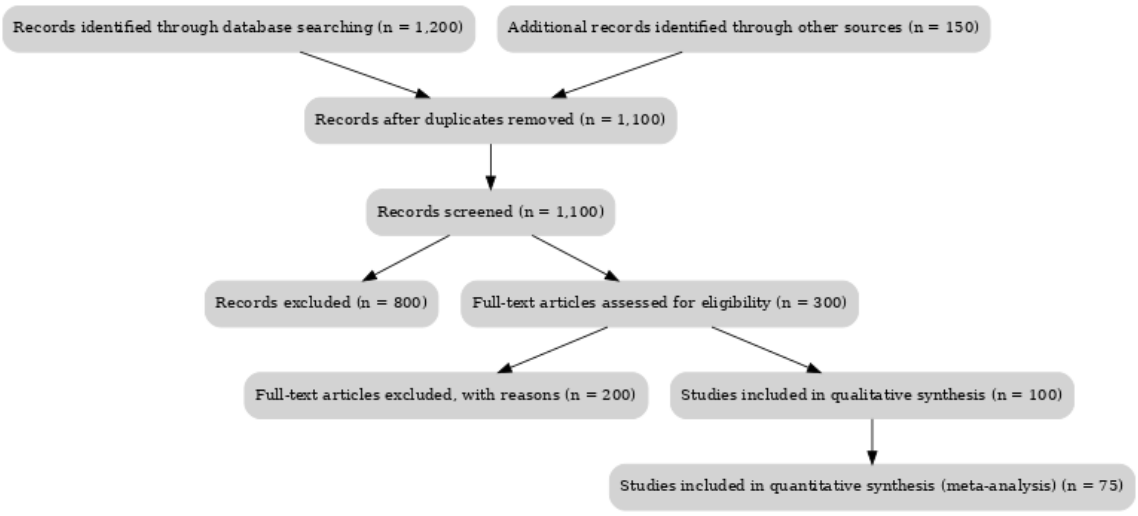

Figure 1: PRISMA flowchart

2.4 Data Extraction And Analysis

A standardized data extraction form was applied to all 75 studies, recording bibliographic metadata, research problem, methodology, datasets, and contributions.

- Bibliographic information: Title, authors, publication year, and IEEE publication venue
- Core problem: The specific challenge or research gap the paper addresses.
- Methodology: The proposed architecture and techniques (e.g., CNN, GCN, Transformer, fusion, etc.).
- Datasets: The benchmark or custom datasets used for evaluation.
- Key contributions: The primary findings and contributions of the work.

A mixed-methods synthesis combined quantitative performance analysis with qualitative coding of methodological innovations. Algorithms were categorized into classical machine learning, deep learning (CNN, RNN, GCN, Transformer), and hybrid/fusion approaches [2]–[6]. Dataset evaluation considered scale, modality, annotation quality, and demographic coverage [7]–[12].

3. Summary of key observations

This section provides an overview of the key findings from the present study, which include the structure of the HAR system architecture, application areas, and datasets used, that serve as the foundation for evaluation and benchmarking.

3.1 The HAR system architecture

The fundamental purpose of a Human Action Recognition (HAR) system architecture is to provide a structured, multi-stage pipeline that methodically transforms raw, noisy sensor data into high-level, interpretable action labels [1]. Its importance lies in establishing a standardized process that ensures modularity and reliability. Each stage from data acquisition to final classification serves as a distinct processing block, allowing researchers to innovate on specific components, such as feature extraction or classification algorithms, while maintaining a coherent end-to-end workflow. This systematic approach is crucial for developing robust and accurate systems capable of functioning in complex, real-world environments [11].

A HAR system typically follows a modular pipeline that transforms raw sensor inputs into

interpretable action labels [1]. The pipeline consists of four major stages: data acquisition, preprocessing, feature extraction, and classification. This structure ensures modularity, enabling innovation at each stage while maintaining an end-to-end framework [2].

At the acquisition stage, systems employ multiple sensing modalities including RGB cameras, depth sensors, inertial measurement units (IMUs), infrared sensors, and even Wi-Fi-based radio frequency devices [3], [4]. The preprocessing stage addresses noise removal, normalization, and temporal alignment.

The feature extraction stage has undergone the most rapid evolution. Traditional handcrafted features have been replaced by deep learning models. Convolutional Neural Networks (CNNs) dominate image-based feature extraction, while Graph Convolutional Networks (GCNs) are widely adopted for skeleton-based recognition due to their ability to capture spatial–temporal topologies [5], [6]. More recently, Transformers have emerged as powerful architectures for multimodal fusion and long-range temporal modeling [7], [8].

Several architectural innovations highlight this progress. Pose-guided GCNs (PG-GCN) improve robustness by integrating body joint relationships into graph learning [6]. Lightweight models such as GNet-FHO reduce computational complexity while retaining competitive accuracy [5]. These efforts address real-world challenges, including deployment in mobile and embedded platforms.

Despite these advances, two persistent challenges remain: (i) the interpretability of deep architectures and (ii) the reliance on curated datasets, which may not reflect uncontrolled environments [9]. Future HAR systems are expected to employ end-to-end trainable architectures with dynamic feedback between modules and built-in mechanisms for privacy preservation, uncertainty modeling, and domain generalization [3], [10].

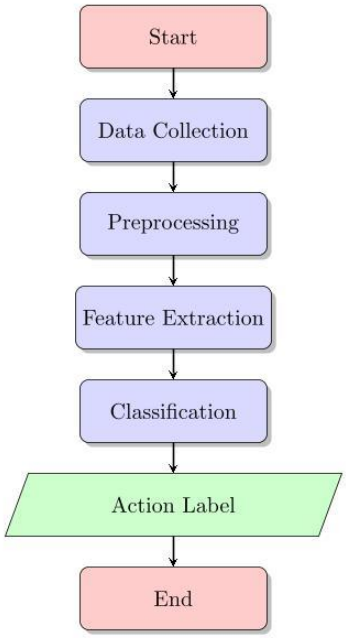

Figure 2: Architectural overview of the HAR system, showing sequential processing stages from sensor data collection to activity labeling.

3.2 Application areas

The translation of HAR research into practical domains demonstrates its societal value. The primary application areas include healthcare, sports analytics, and human–robot interaction (HRI).

In healthcare, HAR enables diagnostic and monitoring systems. For example, deep learning models using pose estimation have been applied to quantify motor impairment in Parkinson's disease [11], while sensor- and vision-based models support early stroke detection [12]. In elderly care, datasets such as HDIA capture daily activities to monitor safety and independence [3].

In sports, HAR systems provide fine-grained analysis of athletic performance. Architectures such as DDC3N and GCN-based frameworks have been used to study tennis serves, CrossFit movements, and figure skating sequences [13], [14]. These models not only enhance training outcomes but also support injury prevention through biomechanical assessment.

In HRI, HAR is a critical enabler for collaborative robots (cobots). By integrating game theory and fuzzy logic, models predict human intent, allowing robots to adjust behavior dynamically [15]. This area has profound implications for industrial safety, interactive gaming, and assistive robotics.

Key challenges across these domains include the scarcity of domain-specific datasets, difficulties in modeling subtle movements (e.g., fine motor impairments), and the need for explainable AI (XAI) to ensure trust and transparency [16].

3.3 Datasets

Datasets are the foundation of HAR research, providing both training material and standardized benchmarks. They fall into two broad categories: benchmark datasets and domain-specific datasets.

3.3.1 Benchmark datasets

Benchmark datasets play a vital role in HAR research because they serve as a fundamental foundation for the development and assessment of models, as illustrated in Table 1. They provide controlled settings for testing and comparing algorithms.

- **NTU RGB+D (60/120 classes):** Large-scale multimodal dataset including RGB, depth, skeleton, and infrared modalities. It is the most widely used benchmark for skeleton-based recognition [6], [17]. Its limitations include controlled indoor settings and noise in skeleton sequences.
- **Kinetics (400/600/700):** A massive video dataset from YouTube clips, serving as the standard for pretraining large-scale deep networks [13]. Challenges include dataset decay (video unavailability) and label noise.
- **UCF101 & HMDB51:** Classical video benchmarks used for generalization studies. Although smaller and limited in diversity, they remain useful for ablation and efficiency testing [18], [19].
- **Sensor-based datasets (**such as WISDM, PAMAP2, UCI-HAR, etc.**):** Provide accelerometer and gyroscope time-series data for wearable device applications [20], [21], [23], [24], [25], [26]. They are lightweight but lack contextual information, such as object interactions.

Table 1: Summary for benchmark datasets

Dataset	Modality	Frame rate	Resolution	Sample/ Class	Application Scenarios	Tool/Framework	Classifier	Accuracy	Scope of Use
NTU RGB+D (60/120)	RGB, Depth, Skeleton, Infrared	30 FPS (typical)	1920×1080 (RGB, variable)	~56,000 clips / 60 or 120 classes	Indoor activity recognition, general HAR models	PyTorch, OpenPose, Open3D	GCNs, Transformers	Varies (up to 96%)	Benchmark for skeleton-based models; limited by lab setting and skeleton data noise. Useful for pose-based HAR research.
Kinetics (400/600/700)	RGB video	25–30 FPS (varied)	Variable (mostly 480–720p)	~300K+ clips / 400–700 classes	General large-scale action recognition, pre-training	TensorFlow, PyTorch	I3D, SlowFast, ViViT	~70–75% top-1	Massive scale; suitable for pre-training and general tasks, not fine-grained; data decay due to YouTube links.
UCF101	RGB video	25 FPS	320×240	13,320 clips / 101 classes	Legacy benchmark, general action	OpenCV, Caffe, TensorFlow	CNNs, Two-stream network	~85–90%	Good for baseline testing and

					recognition		s		efficiency evaluation; limited diversity; older architecture support.
HMDB51	RGB video	30 FPS	320×240 (low quality)	6,766 clips / 51 classes	Film-based action recognition, robustness testing	MATLAB, Python libraries	SVM, 3D CNNs, LSTM	~60–70%	Used to test robustness on poor quality and diverse scenes; small scale and video noise make it difficult for fine-grained analysis.
Sensor-Based (WISDM, etc.)	Accelerometer, Gyroscope (time-series)	20–100 Hz	N/A	~10,000 – 100,000 + samples / 6–18 classes	Fitness tracking, health monitoring, embedded systems	Scikit-learn, TensorFlow	Decision Trees, LSTMs, CNNs	~85–95%	Lightweight, privacy-preserving use in wearables; lacks visual context, sensitive to placement and user variabil

ity.

3.3.2 User-generated datasets

A significant trend in the reviewed literature is the creation of new, specialized datasets designed to overcome the limitations of existing benchmarks and address specific research questions. These datasets are typically developed in-house by research teams to fill a specific gap. For example, the HADE dataset was created to provide a more diverse set of real-world actions than found in many benchmarks [20]. The HDIA dataset was developed specifically for privacy-preserving elderly care, using IR cameras and wearable sensors to avoid capturing identifiable information [3]. Similarly, the NOL-18 Exercise dataset was created to provide labeled data for the specific task of counting exercise repetitions [16], and the CrossFit/Figure Skating datasets were built to enable fine-grained analysis of complex athletic movements [15]. The primary advantage of user-generated datasets is their high relevance to a specific problem, providing data that is much better suited for training specialized models. However, they are often smaller in scale than large benchmarks and may have inherent biases based on the specific collection environment and participant pool.

3.3.3 Summary comparing HAR datasets

Recent advances in HAR datasets have substantially contributed to the field by enabling the development of more accurate and robust systems. Benchmark and domain-specific datasets have improved in capturing complex human actions, interactions, and subtle variations. However, critical challenges remain. Dataset diversity, representativeness, and real-world complexity are often insufficient, limiting the generalization of HAR models across different populations, environments, and sensor types [3], [20]. Future dataset development should prioritize:

1. Inclusivity: Incorporating a broad range of participants, demographics, and body types.
2. Real-world complexity: Including occlusions, multi-person interactions, variable lighting, and diverse environments.
3. Privacy-preserving data acquisition: Utilizing methods such as silhouette representations, anonymized skeletons, or edge-based data processing to minimize sensitive information exposure [22], [23].

These improvements are expected to enhance the universality, precision, and adaptability of HAR systems

3.4 Techniques/Algorithms

HAR research has evolved from classical machine learning methods to advanced deep learning architectures, as illustrated in Table 2. Algorithmic progress can be categorized as follows:

3.4.1 Supervised learning:

Supervised learning remains the dominant paradigm in HAR, where models are trained on labeled datasets. Traditional classifiers, such as Support Vector Machines (SVM) and Random Forests (RF), often perform robustly on sensor-based datasets with handcrafted features. For example, [4] demonstrated that classical models can outperform deep learning methods in cross-subject scenarios due to better generalization. In medical applications, [8] used Logistic Regression and Decision Trees for stroke detection from neuroimaging.

3.4.2 Human–Robot Interaction (HRI)

HRI tasks require algorithms that model human intent and ensure safe collaboration. [6] introduced a cobot decision-making framework that combines game theory with intuitionistic fuzzy sets, allowing robots to account for human hesitation and subjective risk perception. This approach is critical for shared workspaces, enhancing both safety and human-like behavior.

3.4.3 Silhouette sequences

Silhouette sequences are commonly used due to their computational efficiency and privacy preservation. [22] proposed Polygon Coding, a method that converts 2D silhouettes into polygonal representations and encodes geometric properties into fixed-length feature vectors. This eliminates variable sequence length issues without relying on recurrent architectures.

3.4.4 Computational modeling

Efficient computation is essential for edge deployment. Lightweight architectures, such as GNet-FHO, employ Ghost Networks and Fire-Hawk Optimizers to optimize feature selection on wearable sensors [5]. Similarly, Lightweight Video Vision Transformers (LWV-ViT) use spatial-temporal pruning and cross-temporal token interactions for efficient video recognition on edge devices [27].

3.4.5 Graph-based approaches

Graph Convolutional Networks (GCNs) are well-suited for skeleton-based HAR due to their ability to model spatial–temporal joint dependencies. [7] introduced Pose-Guided GCN (PG-GCN), which integrates 2D pose information with 3D skeletons via dynamic attention. In domain-specific contexts, [14] applied GCNs to analyze tennis movements with high precision.

3.4.6 Deep learning architectures

The vast majority of modern HAR systems are built on various deep learning architectures.

- CNNs and RNNs: Hybrid CNN-RNN models, such as Multichannel CNN-GRU [28] and CNN-LSTM with Self-Attention [25], extract spatial features and model temporal dependencies from sensor data.
- Transformers: Self-attention and cross-attention mechanisms excel in multimodal fusion. The SSRT model [2] fuses skeleton and RGB data for fine-grained human-object interaction (HOI), while [18] employs Transformers to learn 3D skeletal representations directly from meshes.
- Specialized Networks: Examples include ResNet-SE for complex activity recognition from wearable sensors [26] and DDC3N, a Doppler-driven 3D CNN for high-precision sports analytics [15].

3.5 Open challenges and limitations

Despite significant progress, the field of HAR systems faces multiple challenges.

3.5.1 Data collection and pre-processing

Collecting and annotating datasets are time-consuming, costly, and prone to noise. Sensor and imaging data are affected by device variability, environmental factors, and acquisition conditions [23], [29]. Preprocessing, such as filtering and normalization, is essential to create clean input for learning algorithms [24].

3.5.2 Dataset modeling and generalization

HAR models often suffer from domain shift, performing poorly on unseen subjects or environments. [4] demonstrated substantial cross-subject performance drops. Large-scale, diverse datasets like HADE [20] and HDIA [3] aim to mitigate this by providing varied training examples.

Table 2: Summary of literature on HAR techniques

Authors	Contributions
[6]	Proposes a cobot action decision-making method based on intuitionistic fuzzy sets and game theory for HRC.
[4],[32]	Compares cross-subject performance of traditional ML and deep learning models on HAR datasets.
[30]	Develops an angular features-based HAR system for real-world applications with subtle unit actions.
[18]	Learns a 3D skeletal representation from a Transformer architecture for action recognition.
[28]	Proposes a multichannel CNN-GRU model for sensor-based human activity recognition.
[26]	Develops a ResNet-SE channel attention-based deep residual network for complex activity recognition.
[10]	Creates an automatic detection pipeline for assessing the motor severity of Parkinson's disease.
[25]	Proposes a deep CNN-LSTM with a self-attention model for HAR using wearable sensors.
[13]	Proposes a framework for learning spatial affordances from 3D point clouds to map unseen human actions.
[8]	Develops a machine learning-based diagnostic model using neuroimages for stroke identification.
[15]	Proposes a Doppler-Driven 3D CNN (DDC3N) for HAR, with new datasets for CrossFit and Figure Skating.
[17]	Provides a comprehensive survey of RGB-based and skeleton-based HAR methods.
[24]	Proposes a multi-stream TCN-based approach with ECA-Net for sensor-based HAR.

3.5.3 Open-access and commercial tools

Open-source tools like OpenPose provide flexible pose estimation pipelines [11], while commercial sensors, e.g., Microsoft Kinect, enabled widespread skeleton-based HAR [9]. Researchers must balance cost, flexibility, and technical complexity when choosing tools.

3.5.4 Video frame analysis

Processing video data is computationally intensive. Techniques such as frame sampling or action segmentation are employed to reduce redundancy. For instance, [16] segmented repetitive exercises into unit actions, while multimodal fusion (RGB + skeleton) resolves ambiguities in human-object interactions [2].

3.5.5 Performance metrics

Accuracy alone can be misleading, particularly in class-imbalanced or safety-critical applications. A combination of precision, recall, F1-score, and task-specific metrics (e.g., ROC, AUC for medical diagnosis) is increasingly recommended [5], [8] as illustrated in Table 3.

Table 3: Overview of datasets and evaluation metrics in recent HAR studies

Authors	Datasets	Evaluation Metrics
[8]	Custom CT Image Dataset	Accuracy, Precision, Recall, F1-score, ROC, AUC
[23]	mHealth, PAMAP2, UCIDSADS	Accuracy, F1-score, Confusion Matrix
[5]	WISDM, Motion Sense, UCI-HAR	Accuracy, Precision, Recall, F1-score
[12]	StanWiFi, MultiEnvironment	Accuracy, Precision, Recall, F1-score
[2]	Toyota Smarthome, ETRI-Activity3D	Accuracy, Precision, Recall, F1-score

In specialized applications like stroke detection, even more advanced metrics like the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are used to assess the diagnostic power of a model across different thresholds [8].

Table 4: HAR taxonomy

S/N	Techniques	Application	Explanation
1.	Computational Modeling	Dynamic	Real-time FPGA-based devices can recognize human actions. Intelligent settings, human-machine communications, and security systems utilize this technology.
2.	Silhouette Sequence Point Clouds	Dynamic	This approach analyzes the time sequence of the camera silhouettes. They have built action-based spaces. The activities and shape information were recognized using 3-D point clouds.
3.	Graph-based approach	Static	Classification of human behavior based on graphs. This model maintains a complex spatial arrangement of the joints in the body by considering how they move and change over time.
4.	Human motion understanding for HRI [32]	Dynamic	New hardware for action recognition based on two-stream neural networks. This design delivers the same accuracy as existing baseline models with fewer operations.
5.	Deep Learning Architectures	Dynamic	This includes various architectures like CNN-LSTM, GCNs, and Transformers to capture complex spatial and temporal patterns in HAR data.
6.	Pre-trained CNNs	Dynamic	Combines class-based and instance-based success rates to assess transfer models. All class- and instance-based NASNet-Large parameterize the ABC-optimized CNN.

4. Highlighted open problems and research gaps in HAR

This section synthesizes current challenges and critical gaps remaining within the HAR field. Table 4 provides a taxonomy of prominent HAR techniques, differentiating them by their underlying computational modeling, application type (static/dynamic), and mechanism, thereby underscoring the diversity and complexity of the current solution landscape. Despite the advancements outlined in this taxonomy, fundamental limitations persist, particularly concerning data heterogeneity and labeling, robustness and generalization, and real-time resource efficiency, which are discussed in detail in the following subsections.

4.1 Cross-subject and cross-dataset generalization

Despite the growth of large-scale HAR datasets, models still struggle to generalize across unseen subjects, environments, or devices. Current deep learning systems often overfit training datasets, limiting real-world deployment. Future research should focus on domain generalization, federated learning, and subject-invariant feature extraction to create models that are robust to inter-subject variability and adaptable to new contexts [4][20]. Developing standardized benchmarks specifically

targeting cross-subject evaluation could further guide the community in addressing this challenge.

4.2 Multimodal fusion for heterogeneous inputs

While multimodal fusion is increasingly adopted, most existing systems rely on simple concatenation or late-fusion strategies. There is a critical need for dynamic, context-aware fusion techniques, such as cross-attention or co-attention mechanisms, which allow different modalities (RGB, skeleton, sensor, audio) to influence each other throughout the pipeline [2][18]. Research in flexible, transformer-based architectures that can handle a variable number of modalities, rather than being fixed, is still an open area that could significantly enhance recognition performance and robustness.

4.3 Data scarcity and synthetic data generation

Many specialized HAR applications, such as healthcare monitoring or sports analytics, suffer from insufficient labeled data. Few-shot and zero-shot learning methods are promising but still immature. There is a need for high-fidelity synthetic data pipelines using advanced generative models (GANs, diffusion models) to simulate diverse human actions under variable environmental and sensor conditions [16][17]. Future research should focus on creating automatically labeled, privacy-preserving synthetic datasets to reduce manual data collection effort and enable robust model training for rare or domain-specific actions.

5. Conclusion and future work

This paper has provides a comprehensive analysis of 30 HAR studies published between 2022–2025, highlighting the transition from generalized, single-modality models to specialized, multimodal architectures designed for real-world deployment. Key trends include the dominance of deep learning models (GCNs, Transformers), the growing importance of multimodal fusion, and increasing focus on domain-specific applications in healthcare, sports analytics, and human-robot interaction. Despite progress, significant challenges persist:

1. High computational cost of state-of-the-art models.
2. Scarcity of large-scale, diverse, unbiased datasets.
3. Cross-subject and cross-dataset generalization. Future research directions include:

- Advanced multimodal fusion: Moving beyond concatenation or late-fusion to cross-attention and co-attention models that allow dynamic interaction between RGB, skeleton, sensor, and audio modalities. Developing flexible Transformer architectures capable of handling variable modality inputs can significantly enhance robustness.
- Generalization and fairness: Prioritize cross-subject and cross-dataset generalization using regularization, domain generalization, and federated learning approaches to train models on decentralized data without compromising privacy. Address algorithmic biases related to gender, age, skin tone, and physical ability for equitable deployment.
- Data scarcity: Expand the use of few-shot and zero-shot learning and develop synthetic data pipelines using GANs or diffusion models to generate diverse, automatically labeled datasets for rare or specialized actions.
- Domain-specific real-time applications: Build interactive healthcare systems for rehabilitation

and adaptive cobots in HRI, emphasizing lightweight models for real-time edge deployment without sacrificing accuracy.

In conclusion, advancing multimodal fusion, ensuring fairness, solving data scarcity, and developing efficient domain-specific systems will propel HAR toward creating safer, more intelligent, and adaptable environments.

Funding: No specific funding received for this research.

Data Availability: This work is primarily theoretical, and consequently, neither a novel dataset nor an existing benchmark dataset was utilized to substantiate the reported findings.

Conflicts of Interest: No conflict of interest is stated by the author.

Ehical consideration: Not applicable.

Authors contributions. Conceptualization: USB, SBA; methodology: USB, SBA, validation: USB, SBA; writing—original draft preparation, USB, SBA; writing—review and editing: SBA; visualization: USB, SBA; supervision: SBA; project administration: SBA; The author had approved the final version.

References

[1] Karim, M., Khalid, S., Aleryani, A. et al., (2024). "Human action recognition systems: a review of the trends and state-of-the-art". *IEEE Access*, 12, 36372-36390.

[2] Ghimire, A., Kakani, V., and Kim, H. (2023). "SSRT: A sequential skeleton rgb transformer to recognize fine-grained human-object interactions and action recognition". *IEEE Access*, 11, 51930-51948.

[3] Park, J., Ok Yang, K., Park, S. et al., (2025). "Human daily indoor action (hdia) dataset: privacy-preserving human action recognition using infrared camera and wearable armband sensors". *IEEE Access*, 13, 60822-60832.

[4] Yang, Z., Qu, M., Pan, Y. et al., (2022). "Comparing cross-subject performance on human activities recognition using learning models". *IEEE Access*, 10, 95179-95196.

[5] Athota, R. K., and Sumathi, D. (2024). "GNet-FHO: A light weight deep neural network for monitoring human health and activities". *IEEE Access*, 12, 108484-108503.

[6] Liu, B., Fu, W., Wang, W. et al., (2022). "Research on cobot action decision-making method based on intuitionistic fuzzy set and game theory". *IEEE Access*, 10, 103349-103363.

[7] Chen, S., Xu, K., Mi, Z. et al., (2022). "Dual-domain graph convolutional networks for skeleton-based action recognition". *Machine Learning*, 111(7), 2381-2406.

[8] Saleem, M. A., Javeed, A., Akarathanawat, W. et al., (2024). "Innovations in stroke identification: a machine learning-based diagnostic model using neuroimages". *IEEE Access*, 12, 35754-35764.

[9] Fang, X., and Guo, Y. (2024). "Human animation model generation in traffic accident restoration: human action recognition based on improved dtw algorithm". *IEEE Access*, 12, 107570-107582.

[10] Yang, N., Liu, D., Liu, T. et al., (2022). "Automatic detection pipeline for accessing the motor severity of parkinson's disease in finger tapping and postural stability". *IEEE Access*, 10, 66961-66973.

[11] Gupta, C., Gill, N. S., Gulia, P. et al., (2024). "A real-time 3-dimensional object detection based human action recognition model". *IEEE Open Journal of the Computer Society*, 5, 14-26.

[12] Jannat, M. K. A., Islam, M. S., Yang, S. et al., (2023). "Efficient Wi-Fi-based human activity recognition using adaptive antenna elimination". *IEEE Access*, 11, 105440-105454.

[13] Piyathilaka, L., Kodagoda, S., Thiyagarajan, K. et al., (2024). "Learning spatial affordances from 3d point clouds for mapping unseen human actions in indoor environments". *IEEE Access*, 12, 868-877.

[14] Zhang, X., and Chen, J. (2023). "A tennis training action analysis model based on graph convolutional neural network". *IEEE Access*, 11, 113264-113271.

[15] Toshpulatov, M., Lee, W., Lee, S. et al., (2024). "DDC3N: Doppler-driven convolutional 3d network for human action recognition". *IEEE Access*, 12, 93546-93567.

[16] Cheng, S., Sarwar, M. A., Daraghmi, Y. et al., (2023). "Periodic physical activity information segmentation, counting and recognition from video". *IEEE Access*, 11, 23019-23031.

[17] Wang, C., and Yan, J. (2023). "A comprehensive survey of rgb-based and skeleton-based human action recognition". *IEEE Access*, 11, 53880-53898.

[18] Cha, J., Saqlain, M., Kim, D. et al., (2022). "Learning 3d skeletal representation from transformer for action recognition". *IEEE Access*, 10, 67541-67550.

[19] Xu, Q., Yang, J., Zhang, H. et al., (2024). "Enhancing few-shot action recognition using skeleton temporal

alignment and adversarial training". *IEEE Access*, 12, 31745-31755.

[20] Karim, M., Khalid, S., Aleryani, A. et al., (2024). "HADE: exploiting human action recognition through fine-tuned deep learning methods". *IEEE Access*, 12, 42769-42790.

[21] Huang, K., Mckeever, S., and Miralles-Pechuán, L. (2024). "Generalized zero-shot learning for action recognition fusing text and image GANs". *IEEE Access*, 12, 5188-5202.

[22] Göçmen, O., and Akata, M. E. (2023). "Polygonized silhouettes and polygon coding based feature representation for human action recognition". *IEEE Access*, 11, 57021-57036.

[23] Sharif, U., Mehmood, Z., Mahmood, T., Javid, M. A., et al., (2019). "Scene analysis and search using local features and support vector machine for effective content-based image retrieval". *Artificial Intelligence Review*, 52, 901-925. https://doi.org/10.1007/s10462-018-9636-0

[24] Miah, A. S. M., Hwang, Y. S., and Shin, J. (2024). "Sensor-based human activity recognition based on multi-stream time-varying features with ECA-Net dimensionality reduction". *IEEE Access*, 12, 151649-151668.

[25] Khatun, M. A., Yousuf, M. A., Ahmed, S. et al., (2022). "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor". *IEEE Journal of Translational Engineering in health and Medicine*, 10, 1-16.

[26] Mekruksavanich, S., Jitpattanakul, A., Sitthithakerngkiet, K. et al., (2022). "ResNet-SE: channel attention-based deep residual network for complex activity recognition using wrist-worn wearable sensors". *IEEE Access*, 10, 51142-51154.

[27] Han, J., Zhao, J., Yue, Y. et al., (2024). "Edge computing-based video action recognition method and its application in online physical education teaching". *IEEE Access*, 12, 148666-148676.

[28] Lu, L., Zhang, C., Cao, K. et al., (2022). "A multichannel CNN-GRU model for human activity recognition". *IEEE Access*, 10, 66797-66810.

[29] Man, K., Chahl, J., Mayer, W. et al., (2024). "The effects of different image parameters on human action recognition models trained on real and synthetic image data". *IEEE Access*, 12, 95223-95244.

[30] Ryu, J., Patil, A. K., Chakravarthi, B. et al., (2022). "Angular features-based human action recognition system for a real application with subtle unit actions". *IEEE Access*, 10, 9645-9657.

[31] Abdullahi, S.B., and Chamnoongthai, K. (2022). "American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach". *IEEE Access*, 10, 15911-15923.

[32] Abdullahi, S.B., and Chamnoongthai, K. (2023). "IDF-sign: Addressing inconsistent depth features for dynamic sign word recognition". *IEEE Access*, 11, 88511-88526.

[33] Alamri, F.S., Abdullahi, S.B., Rehman, A.K., and Saba, T. (2024). "Enhanced weak spatial modeling through CNN-based deep sign language skeletal feature transformation". *IEEE Access*, 12, 77019-77040.

Appendix A

LIST OF ABBREVIATIONS

Abbreviation	Full Name
HAR	Human Action Recognition
HOI	Human-Object Interaction
ML	Machine Learning
DL	Deep Learning
GCN	Graph Convolutional Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
IMU	Inertial Measurement Unit
IR	Infrared
CSI	Channel State Information
DTW	Dynamic Time Warping
S-SVM	Structured Support Vector Machine
GZSAR	Generalized Zero-Shot Action Recognition
LOSO	Leave-One-Subject-Out
PD	Parkinson's Disease
HRC	Human-Robot Collaboration