

Review

Computer Vision Models for Human Activity Recognition: A Literature Review

Luís Henrique Travassos ^{1,*}, Edite Ravella ², Jorge Bernardino ^{1,3,*} and Francisco B. Pereira ^{1,3}

¹ Department of Informatics and Systems Engineering, Coimbra Institute of Engineering, Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal; xico@isec.pt

² SafeCaring, Rua Pedro Nunes, s/n, 3030-199 Coimbra, Portugal; edite@safecaring.pt

³ CISUC/LASI—Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, P.l.o II—Pinhal de Marrocos, 3030-290 Coimbra, Portugal

* Correspondence: a2021136600@isec.pt (L.H.T.); jorge@isec.pt (J.B.)

Abstract

Human Activity Recognition (HAR) is the automated process of identifying human actions using sensor data or video, which is widely used in healthcare, smart environments, and surveillance. Although HAR based on computer vision has advanced rapidly, existing reviews do not adequately address the recent shift toward hybrid deep-learning architectures or provide a structured comparison of the trade-offs relevant to real-world deployment. This literature review addresses that gap through a PRISMA-guided analysis of articles published between 2021 and 2025 and retrieved from four major databases. The review develops a reproducible taxonomy of nine architectural families and applies a multidimensional evaluation framework covering classification accuracy, computational efficiency for edge deployment, environmental generalization, and fine-grained activity recognition. The findings show that hybrid architectures are the dominant design strategy, while attention-based and graph-based models play important specialized roles depending on temporal complexity, privacy requirements, and deployment constraints, with the literature concentrated mainly in healthcare and security applications.

Keywords: computer vision; deep learning; human activity recognition; literature review; performance evaluation

1. Introduction

Human Activity Recognition (HAR) is a technology that has gained significant awareness in the field of intelligent system development by enabling machines to see, understand, and interpret what a human is doing based on the data gathered from those actions [1,2].

Historically, early HAR models gathered their data mainly from wearable sensors [3], such as smartphones with gyroscopes, accelerometers, compasses, and others, as well as bracelets, smart bands, and other wearable sensors. However, advancements in hardware and software components, such as cameras and computer vision, have made vision-based approaches dominant and promising [4,5]. These methods use detailed information on location and time in a video to identify activities such as walking, running, jumping, and falling. Unlike sensor-based methods, which are physically intrusive, vision-based methods are non-contact and can be used on a large scale [6], providing a better understanding of specific human actions in the places where they happen.



Academic Editor: Guanqiu Qi

Received: 29 April 2026

Revised: 16 June 2026

Accepted: 18 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

The field of HAR is changing rapidly due to the evolution of deep learning models. Architectural patterns have progressed from manual feature extraction and basic classifiers to advanced hybrid architectures that combine complementary mechanisms for spatial, temporal, and contextual modelling [2,7].

However, the review literature has not kept pace with this transition, particularly in systematically organizing the growing diversity of hybrid and specialized architectures that now define vision-based HAR. This gap creates a practical challenge for both researchers and practitioners. Without an updated synthesis, researchers lack a coherent frame for positioning new contributions and situating them within current architectural trends. At the same time, practitioners in domains such as healthcare, surveillance, and industrial monitoring are often forced to rely on isolated benchmark results when selecting models, rather than on a structured understanding of the trade-offs among accuracy, efficiency, robustness, privacy, and interpretability. In the absence of such a review, model selection would remain fragmented, less transparent, and more difficult to justify in real-world deployment settings.

The relevance of this challenge is increasingly evident in industry-driven systems that aim to translate advances in academic research into practice. SafeCaring [8], for example, is developing AI-powered computer vision systems for patient rooms designed to automate clinical documentation and improve patient safety. Applications of this kind require more than high predictive accuracy; they require architectures that are reliable, efficient, interpretable, privacy-aware, and capable of operating effectively in constrained, high-stakes environments.

Building on this context, this review provides a current and structured account of the vision-based HAR literature published since 2021. Using a PRISMA-guided analysis, the review classifies the main computer vision model families used in HAR, analyzes the trade-offs among them, and identifies the application domains in which they are most prominently adopted. Rather than serving only as a descriptive survey, the review offers a decision-oriented synthesis by developing a taxonomy of vision-based HAR architectures, extending comparison beyond raw accuracy to include computational efficiency, environmental generalization, and fine-grained activity discrimination, and mapping architectural choices to field-specific needs. In doing so, the review clarifies the current state of knowledge, highlights areas where the literature remains fragmented, and examines how contemporary architectural strategies align with real-world deployment demands.

The paper is organized as follows: Section 2 establishes the fundamentals of computer vision for activity recognition, detailing the processing pipeline and the current technological challenges. Section 3 describes the review methodology, including the PRISMA-based search strategy and selection criteria. Section 4 presents the thematic analysis and findings addressing model architectures, performance comparisons, and application contexts. Section 5 offers a critical discussion of current trends, reflecting on the field's strategic evolution, and suggesting future research directions. Finally, Section 6 provides the concluding remarks, highlighting the existing research gaps and synthesising the primary findings.

2. Fundamentals of Computer Vision for Activity Recognition

HAR is a multidisciplinary field at the intersection of computer vision, pattern recognition, and machine learning that is focused on the automated interpretation of human activities from observational data. HAR primary goal is to enable machines to understand and categorize human actions, behaviours, and interactions within a given environment [1,9,10]. Early HAR systems relied heavily on data from wearable sensors, such as accelerometers and gyroscopes, that for themselves, relied heavily on manual work for feature engineering. However, the advent of powerful computer vision techniques and the proliferation of

visual data retrieval cameras have positioned vision-based approaches as a dominant and highly promising new paradigm [2,11].

Vision-based HAR analyses the spatio-temporal dynamics of human movement using sequential images or video data [3,12]. The fundamental processing pipeline is illustrated in Figure 1 demonstrating the steps involved in vision-based HAR model.

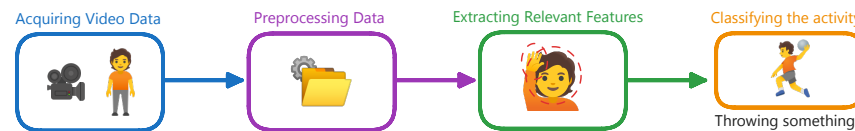


Figure 1. General Pipeline for Vision-Based HAR.

As shown, it typically involves acquiring raw video data, pre-processing it to enhance quality and consistency, then extracting discriminative features that capture relevant motion and shape patterns, and at last, classifying the sequence into a specific activity category [13,14].

The evolution of feature extraction has been particularly transformative. Initial methods depended on handcrafted features, such as histograms of oriented gradients and optical flow, which were designed to explicitly capture human shape and motion patterns. While these methods were effective in constrained settings, they often struggled with variability in viewpoint, lighting, clothing, and occlusions [15–17].

Breakthroughs in modern HAR have been driven by deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Early foundational works such as 3D Convolutional Networks (C3D) and the Two-Stream CNN demonstrated that understanding videos improves when spatial appearance and temporal motion are learned either jointly from video volumes or through separate RGB and optical-flow streams [18,19].

CNN models excel at extracting robust spatial features from individual frames and learning hierarchical representations of body postures and immediate context. However, activities are inherently temporal. To model this dimension, CNNs are often combined with architectures such as Long Short-Term Memory (LSTM) networks or Gated Recurrent Units (GRU) to form spatio-temporal models that understand how a sequence of postures evolves into an action [20–22].

A significant trend in the field is the shift towards using skeleton-based HAR. Rather than processing raw RGB pixel data, these methods use pose estimation models to extract a simplified, high-level representation of the human body as a set of key points connected by lines. This skeleton data is efficient, privacy-preserving, and invariant to superficial appearance changes, making it highly suitable for applications requiring robustness and deployment on resource-constrained devices [10,23].

The classification task then shifts from extracting and analysing simple motion and shape patterns, to analysing the spatio-temporal evolution of the information gathered from the skeletal models and graphs. This direction was strongly shaped by the Spatial Temporal Graph Convolutional Network (ST-GCN), which formalized skeleton sequences as graphs connected across both body structure and time and became a foundational architecture for later graph-based HAR models [24–26].

Despite these advancements, HAR continues to face several core challenges. One of the primary limitations is the dependency for large-scale, well-annotated datasets required for training robust models, which are often costly and time-consuming. In addition, HAR systems struggle with generalization across different viewpoints and environmental conditions. There is also the difficulty of recognizing fine-grained activities (for example, “drink-

ing water” versus “brushing teeth”), and models that rely on computationally intensive architectures that hinder the deployment in real-time settings or on edge devices [27–29].

3. Review Methodology

This study adhered to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [30], comprising the following sequential phases of identification, screening, eligibility, and inclusion. Each of these phases are detailed in the following subsections and summarized in Figure 2.

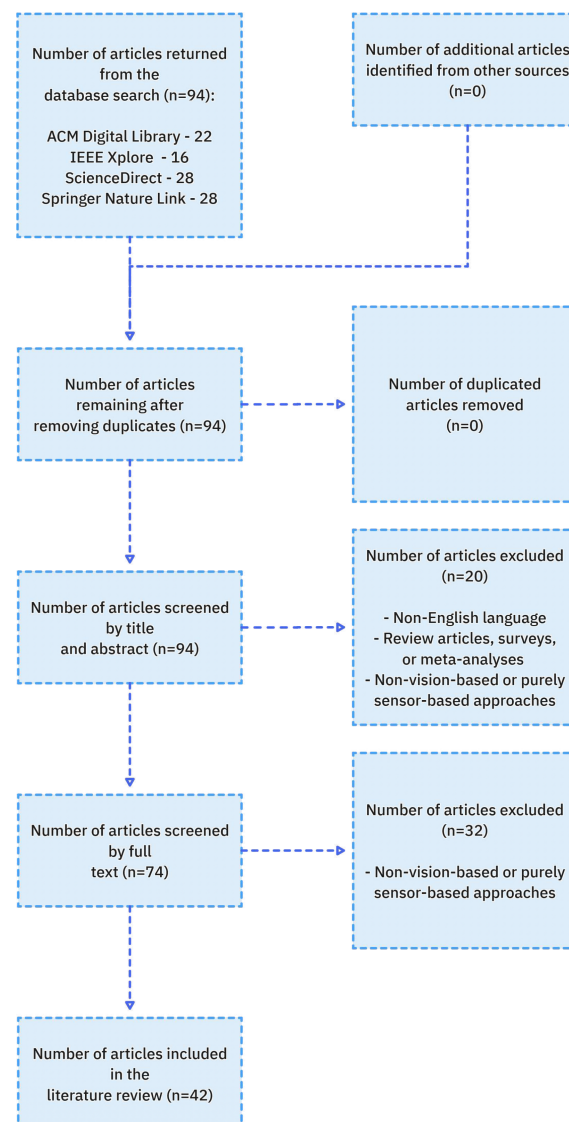


Figure 2. PRISMA Flow Diagram.

3.1. Search Strategy

The initial phase consisted of a comprehensive search across four prominent databases namely, the ACM Digital Library, IEEE Xplore, ScienceDirect, and Springer Nature Link. The eligibility criteria for the selection of these databases was based on their comprehensive and general coverage on the leading research and developments in the field of computer science and engineering, between 1 January 2021, and 31 December 2025.

To guarantee the inclusion of the most recent advancements, the search was strategically confined to articles published between 1 January 2021, until 31 December 2025.

The query was adapted to each database's syntax while retaining its meaning, capturing advances in computer vision and visual generative domains and focusing on primary studies. This was achieved by combining "computer vision" (the ability of a computer to recognize and interpret what it sees) or "visual generative" (the creation of new images) with "model" and "activity recognition", as the term "model" is used to focus on innovation and technical development. The core generic search string used was as follows:

"computer vision" OR "visual generative" AND "model" AND "activity recognition"

This query ensured the retrieval of articles that focus explicitly on computer vision models. The initial database search returned 94 unique articles. No additional records were identified from other sources.

3.2. Exclusion Criteria

A two-stage screening process was applied to the 94 articles that were identified to determine their eligibility. In the first stage, titles and abstracts were screened against predefined exclusion criteria. Articles were excluded if they were written in a non-English language, corresponded to review articles, surveys, or meta-analyses, or primarily focused on non-vision-based approaches, such as embedded sensors, frequency sensors, inertial measurement unit sensors, smartphone sensors, and wearable sensors.

This stage resulted in the exclusion of 20 articles. The remaining 74 articles underwent a full-text assessment for eligibility, resulting in the exclusion of an additional 32 articles for reasons consistent with the initial criteria. The final set of 42 articles included in the review underwent the following data extraction process.

3.3. Data Extraction

The data extraction process identified the following crucial information in each article: the proposed model architecture; the database used to access the articles; performance metrics; documented challenges; and application context.

Figure 3 illustrates the temporal distribution of the selected literature. Analysing the temporal window of the selected articles, it is clear that 2025 stands out as the most prominent year, accounting for one-third of the total publications with 14 articles included in the review. Another 8 articles are from 2024, 10 articles are from 2023, 7 articles are from 2022, and 3 articles are from 2021. This distribution reflects the field's rapid evolution and ensures the literature review captures recent advancements while maintaining a substantive foundation from preceding years.

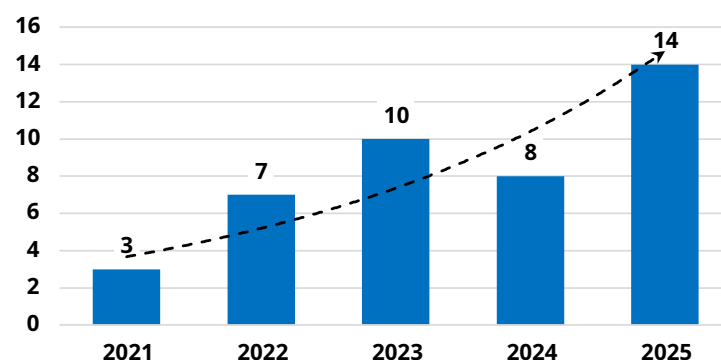


Figure 3. Annual Distribution of Included Publications.

Figure 4 reveals that the articles included were sourced from four primary databases. ScienceDirect contributed with the most publications (16 articles), followed by Springer Nature Link (13 articles), IEEE Xplore (11 articles) and ACM Digital Library (2 articles).

This distribution confirms comprehensive coverage of major publications in the fields of computer science and engineering that are relevant to HAR research.

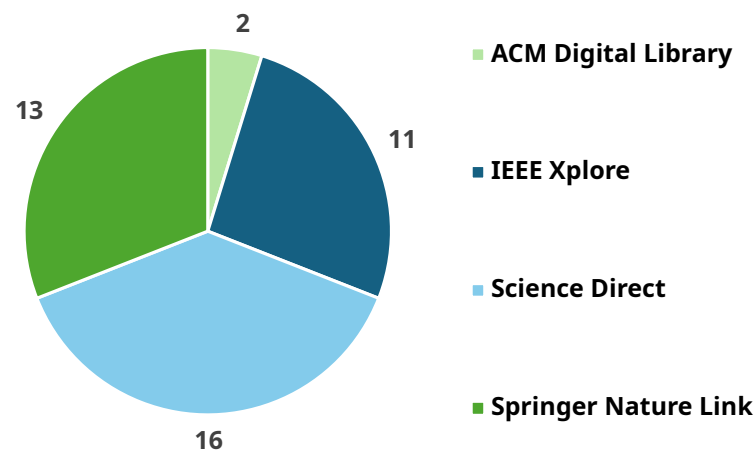


Figure 4. Publication Distribution by Source Database.

After the initial screening and data extraction stages, a qualitative approach was adopted to organize the findings into different categories. Key trends in model development and architectural design were identified, enabling a comparison of how well the different designs performed. Additionally, the use of computer vision-based methods for recognizing human activity was analysed.

4. Thematic Analysis and Findings

This section provides an analysis of the 42 articles chosen for this review and summarizes their contributions to address the defined objectives. To establish a foundation for this thematic discussion, a detailed categorization of the model architectures identified in the literature is presented first. This classification framework provides the structure for the analysis and paves the way to understand the technological landscape of computer vision-based HAR. The analysis revealed nine distinct categories of computational models, many of which use multiple architectural approaches.

These families are the following: Classical Machine Learning, CNN, 3D CNN, Recurrent, Temporal Convolutional Network (TCN), Attention-Based, Graph-Based, Relational, and Hybrid models. Although many articles combine several of these strategies, the taxonomy captures the dominant computational logic behind contemporary vision-based HAR systems and provides a consistent basis for comparison.

A comprehensive overview of the distribution of these categories across the literature corpus is presented in Table 1.

The following subsections first explain how each architectural family operates and why it has been used for HAR, then examine how these families are distributed across the literature, how they perform across multiple evaluation dimensions, and the domains in which they are most frequently applied.

Table 1. Categorization of Reviewed Articles by Application and Architectural Family.

Article	Year	Application	Overview	Model Architectural Family								
				Classical ML Models	CNN Architectures	3D CNN Models	Recurrent Architectures	TCN Models	Attention-Based Models	Graph-Based Models	Relational Models	Hybrid Architectures
[13]	2025	Education	STrack4Re method that uses YOLOv5 and OC-SORT for continuous classroom action detection and tracking.		✓		✓					✓
[31]	2025	Education	It uses 3D-CNN to extract spatio-temporal features from classroom videos, combined with Support Vector Machine (SVM) classifiers.	✓		✓						✓
[17]	2024	Education	It uses YOLOv5 enhanced with motion feature maps to detect global and local anomalies.		✓							
[32]	2025	Generic	It optimizes a MobileNetV2 hybrid model with additional CNN layers to capture spatial and temporal details.		✓							✓
[9]	2025	Generic	It combines actor relationship graphs with GCN-LSTM and attention mechanisms to capture group dynamics.		✓		✓		✓	✓		✓
[10]	2025	Generic	Encodes 3D data from skeletal joints into greyscale images processed by a three-channel fusion network.		✓				✓			✓
[23]	2024	Generic	HAR-ViT method for 3D skeleton data, integrating enhanced Adaptive Graph Convolutional Layer and redefined position encoder.						✓			
[7]	2024	Generic	Model that merges VGG16 (spatial), Bidirectional Gated Recurrent Unit (BiGRU) (temporal), and ResNet18 (high-level) via transfer learning.		✓		✓					✓
[33]	2023	Generic	It focuses on fine-grained human-object interactions using embedded inertial sensors and models such as K-Nearest Neighbors (KNN) and XGBoost.	✓								
[22]	2023	Generic	It replaces CNN and RNN with Vision Transformers (ViT) and Recurrent Transformers to improve speed.						✓			✓
[11]	2022	Generic	It combines CNN (various kernel sizes) and GRU to extract local features and long-term dependencies.		✓		✓					✓
[5]	2022	Generic	Explores the use of transfer learning in DLN architectures to identify gaze fixation areas on screens.		✓		✓					✓
[3]	2022	Generic	Four-branch CNN architecture with attention and feature fusion modules for recognition in still images.		✓				✓			✓
[14]	2022	Generic	Comparative study of supervised and unsupervised models for HAR.		✓		✓					
[16]	2021	Generic	Combines single-channel/Siamese CNN with dynamic time warping distances and a soft voting set.	✓	✓							✓
[27]	2024	Health—Condition Diagnosis	It introduces the Interpretable Feature Reduction Function using 1D-CNN for medical monitoring.		✓							
[26]	2023	Health—Condition Diagnosis	Implements ST-GCN to classify motor activities from the MDS-UPDRS scale in Parkinson’s patients.								✓	
[34]	2023	Health—Condition Diagnosis	Evaluates lightweight feature extractors combined with TCN to detect autism behaviours.		✓		✓	✓				✓
[35]	2025	Health—People Monitoring	It combines ViT for feature extraction with BiGRU-SAM for fall detection.				✓		✓			✓
[36]	2025	Health—People Monitoring	It employs the Deep Belief Network network optimized by Snake Optimizer to classify WISDM activities.	✓								
[1]	2025	Health—People Monitoring	It uses EfficientNetB7 for features and Bidirectional Long Short-Term Memory (BiLSTM) with Multi-Head Self-Attention for classification.		✓		✓		✓			✓
[37]	2023	Health—People Monitoring	Uses joint angles and distances to create a numerical dataset processed by ensemble classifiers.	✓								✓
[15]	2023	Health—People Monitoring	Proposes a multilayer 2D CNN model to interpret inertial sensor sequences from the WISDM database.		✓							
[20]	2022	Health—People Monitoring	It integrates image, object detection, and skeleton features into a single recurrent neural network.		✓		✓					✓
[2]	2025	Health—Physical Procedures	It proposes the DenMobVit hybrid model, fusing CNN and ViT to classify rehabilitation exercises.		✓				✓			✓
[38]	2023	Health—Physical Procedures	Applies random increments in time segments to train CNN+TCN networks in surgeries.		✓			✓				✓

Table 1. Cont.

Article	Year	Application	Overview	Model Architectural Family								
				Classical ML Models	CNN Architectures	3D CNN Models	Recurrent Architectures	TCN Models	Attention-Based Models	Graph-Based Models	Relational Models	Hybrid Architectures
[39]	2025	Industry	It classifies load lifting and lowering movements using OpenPose with Multilayer Perceptron and LSTM models.	✓			✓					
[6]	2025	Industry	Joint learning framework to recognize individual, team, and global actions on construction sites.		✓				✓	✓	✓	✓
[40]	2024	Industry	Implements 2D ConvLSTM and Long-term Recurrent Convolutional Network models to detect distractions in drivers, including diverse profiles (for example, Hijab).		✓		✓					✓
[4]	2022	Industry	It uses hidden Markov models applied to 2D/3D skeleton data to recognize assembly and separation tasks.	✓								
[41]	2023	Sports	Proposes an Attention Relational Network that merges joint vector sequences to identify mutual actions.				✓		✓		✓	✓
[28]	2025	Surveillance	Integrates the multi-scale and attentional features module with bilinear BiGRU to recognize suspicious actions.		✓		✓		✓			✓
[42]	2025	Surveillance	It compares CNN-BiLSTM with a spatio-temporal transformer approach to detect violence.		✓		✓		✓			✓
[21]	2025	Surveillance	Combines ConvNeXt for spatial extraction with TCN or BiLSTM for temporal sequence analysis.		✓		✓	✓	✓		✓	✓
[12]	2024	Surveillance	Framework based on dynamic attention fusion units and spatio-temporal fusion networks.		✓		✓		✓			✓
[43]	2024	Surveillance	It uses a dual stream with ViT and FlowNet2 to extract context and movement, optimized by multiple attention.		✓		✓		✓			✓
[44]	2024	Surveillance	Privacy-preserving framework using Advanced Encryption Standard encryption on Region of Interest detected by YOLOv7 and CNN-LSTM network.		✓		✓					✓
[25]	2023	Surveillance	Hybrid solution (edge-cloud) that utilizes lightweight posture recognition on the device and 3D SlowOnly in the cloud.			✓						✓
[45]	2023	Surveillance	Presents a robust system for detecting suspicious behaviours in institutions (for example, cheating, theft) using YOLOv5 and Mask R-CNN.		✓							
[46]	2022	Surveillance	It proposes the extraction of silhouettes and posture vectors (Golden Ratio) to identify suspicious behaviors in public places.	✓								
[47]	2021	Surveillance	Lightweight framework using LiteFlowNet for pyramidal features and a novel DS-GRU recurrent unit.		✓		✓					✓
[29]	2021	Surveillance	A 63-layer “L4-Branched-ActionNet” model with ACS optimization and SVM/KNN classifiers.	✓	✓							✓

Note : The table above illustrates the comprehensive distribution of model architectures across the 42 included articles, with check marks indicating the utilization of specific model categories in each publication.

4.1. Architectural Families: Definitions and Relevance to HAR

Before comparing the prevalence or performance of different architectural families in HAR, it is necessary to explain the computational role of each family. Human activities combine spatial appearance, temporal evolution, interaction patterns, and practical deployment constraints, which is why different architectural families recur throughout the literature. Table 2 summarizes the operating principle of each family, its specific relevance to HAR, and representative examples drawn from the reviewed corpus.

Table 2. Architecture Families in Vision-Based HAR: Definitions and Representative Examples.

Family	Core Mechanism	Why It Is Used for HAR	Representative Examples from the Reviewed Literature
Classical ML models	Handcrafted descriptors are extracted first, and shallow classifiers such as SVM, Random Forest (RF), or XGBoost make the final prediction.	Useful when activities can be summarized through explicit motion statistics or pose descriptors. Strengths include interpretability and low computational cost. The main limitation is strong dependence on manual feature engineering and poor transfer to visually complex scenes.	SVM, RF, XGBoost [33,37,46]
CNN architectures	Two-dimensional convolutions learn hierarchical visual patterns from image frames, capturing posture, local appearance, and object context.	Suitable for HAR because many activities are visually distinguished by body configuration and scene context. CNNs are strong spatial encoders and transfer well across datasets, but they need auxiliary temporal modules when actions unfold over longer periods.	MobileNetV2, EfficientNetB7, VGG16 [1,7,32]
3D CNN models	Convolutions operate across width, height, and time simultaneously, learning motion and appearance directly from short video clips.	Useful when HAR depends on joint spatio-temporal patterns rather than isolated frames. Their strength is integrated motion modelling. Their main weakness is high computational and data demand.	3D-CNN, 3D SlowOnly [25,31,43]
Recurrent architectures	LSTM, GRU, and ConvLSTM units process frame features sequentially while preserving a memory of previous states.	Appropriate for HAR because many actions are defined by temporal progression, not by a single pose. These models help distinguish activities with similar frames but different orderings, however, sequential execution creates latency and optimization bottlenecks.	LSTM, BiGRU, ConvLSTM [28,40,47]
TCN models	Temporal convolutions with extended receptive fields model sequential dependencies through parallel operations rather than recurrent state updates.	Used in HAR when long sequences must be processed efficiently. TCN retain temporal sensitivity while reducing inference bottlenecks, although performance depends on suitable receptive-field design and temporal granularity.	TCN, CNN+TCN [21,34,38]
Attention-Based models	Self-attention weighs the most informative regions or time steps and can model long-range dependencies through Transformer-style architectures.	Valuable for HAR when relevant cues are distributed across a video and cannot be captured well by only local convolutions. Their strengths are global context modelling and selective focus. Their main limitations are complexity and higher deployment cost.	ViT, Spatio-Temporal Transformer [2,22,23,42]
Graph-Based models	Human joints are represented as nodes and skeletal connections as edges, allowing message passing over body structure.	Effective for HAR because activities can often be described through coordinated joint movement. These models are robust to background clutter and privacy-sensitive settings, but they depend on accurate pose extraction.	ST-GCN, Adaptive Graph Convolutional Layer [23,26]
Relational models	Explicit representations are learned for dependencies between body parts, actors, or groups rather than for isolated individuals only.	Relevant when HAR involves social interaction, collective behavior, or human-object coordination. Their strength is interaction reasoning. Their limitation is that they are specialized and often require richer annotations or more structured inputs.	Attention Relational Network, Actor Relation Graph [9,41]
Hybrid architectures	Two or more families are combined so that complementary modules handle spatial, temporal, structural, or attentional aspects jointly.	Common in HAR because real-world activities rarely depend on a single inductive bias. Hybrid models usually provide the best overall trade-off in accuracy and robustness, but they are harder to design, optimize, and interpret.	DenMobVit, CNN-LSTM, GCN-LSTM-Attention [2,7,9,20]

Note: The families are analytical categories used to organize the literature. Several reviewed systems combine more than one family, especially hybrid approaches that integrate spatial, temporal, structural, or attentional modules.

Classical machine learning models remain relevant when human movement can be represented through joint angles, trajectories, or summary statistics, especially in settings where interpretability, limited training data, and low computational cost are priorities. CNN and 3D CNN families address the spatial and spatio-temporal structure of video more directly: standard CNNs are well suited to HAR because they learn hierarchical visual cues such as posture, silhouette, and object context, while 3D CNN extend this capability across short video volumes to capture motion jointly with appearance.

Recurrent and TCN architectures are motivated by the fact that an activity is defined not only by posture, but by how posture changes over time. Therefore, LSTM-, GRU-, and ConvLSTM-based models remain useful for HAR because their sequential processing preserves temporal memory and helps distinguish actions with similar frames but different temporal order. TCN pursue the same objective through temporal convolutions, offering a more parallel and deployment-friendly solution for processing long sequences efficiently.

The attention-based, graph-based, relational, and hybrid families respond to more specialized HAR demands. Transformer-style self-attention improves the modelling of long-range dependencies and helps isolate the most informative temporal or spatial cues. Graph-based models are particularly suitable for skeleton-based HAR, where structural relationships between joints are central to the task and privacy preservation is often important. Relational models are best suited to multi-person and group activities, where meaning emerges from interactions rather than from individual actors. Hybrid models are especially prevalent because they combine complementary strengths and reflect the practical reality that robust HAR systems usually require spatial, temporal, and contextual reasoning simultaneously.

4.2. Taxonomy and Mapping of Model Architectures

Building on the architectural definitions summarized in Table 2, this subsection examines how the nine identified families are distributed across the reviewed literature. The analysis characterizes their prevalence, functional roles, and architectural patterns, thereby clarifying how the conceptual taxonomy is reflected in published HAR research.

Figure 5 provides a quantitative overview of the prevalence of the nine main model categories identified in the literature, sorted by architectural complexity, highlighting clear preferences in model selection and revealing trends in the evolution of architectural approaches within the HAR research landscape.

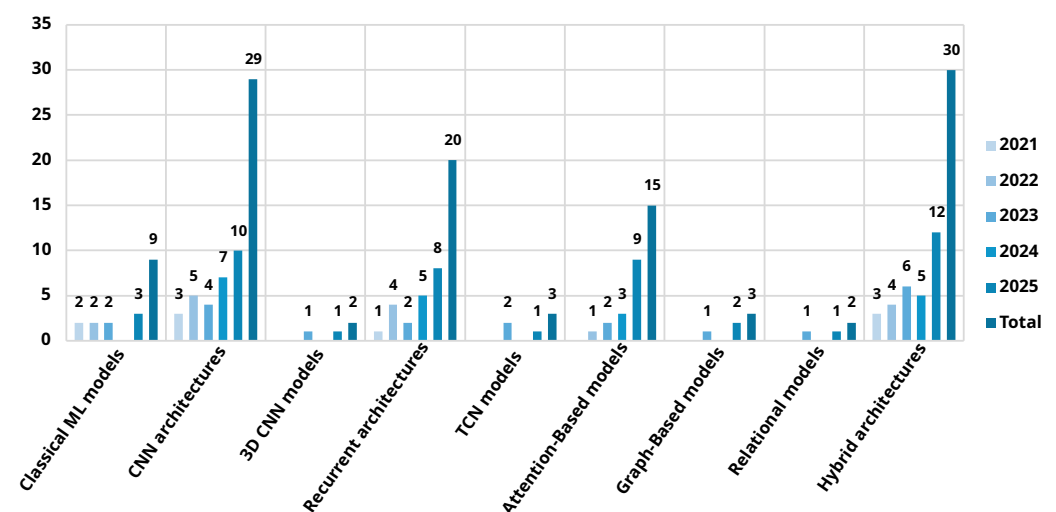


Figure 5. Distribution of Model Architectures Across Reviewed Literature.

In the reviewed corpus, hybrid labels co-occur in 71% articles, yet CNN (69%) and recurrent modules (48%) remain nearly as frequent as building blocks, suggesting fusion rather than replacement of single-family designs. This pattern indicates a recurring design choice when accuracy and robustness are prioritised over minimal complexity, with models typically integrating spatial feature extractors and temporal or attention processors.

CNN architectures appear in 69% of articles, while recurrent architectures appear in 48%. Well-established backbones such as ResNet, VGG, and MobileNet function primarily as spatial encoders within broader pipelines, and recurrent networks such as LSTM and GRU frequently serve as the sequential component in hybrid systems rather than as standalone solutions.

Attention-based models are referenced to in 36% of articles, representing a significant and growing trend. Their adoption reflects a broader shift toward self-attention mechanisms and Transformer-based architectures for modelling long-range dependencies and dynamically emphasizing salient spatial and temporal cues. ViT have emerged as a prominent architecture in this category.

The remaining categories demonstrate more specialized applications. Classical machine learning models, including SVM and RF, are discussed in 21% of articles, often serving as baseline classifiers or in scenarios with resource constraints. TCN and graph-based models appear in 7% of articles, each. TCN offer an alternative to RNN for temporal modelling, while graph-based approaches excel in skeleton-based recognition.

3D CNN models (5% of articles) provide integrated spatio-temporal processing. Relational models (5% of articles) are the most specialized category, reserved for recognizing complex group activities that require explicit modelling of interpersonal relationships.

The taxonomic analysis reveals a three-part structure in the reviewed corpus:

- Dominant use of Hybrid architectures that strategically combine CNN-based spatial feature extractors with Recurrent or Attention-based temporal models.
- Growing use of Attention-Based models, particularly ViT, for capturing long-range spatio-temporal dependencies.
- A set of specialized models, including Graph-Based networks for skeleton data and 3D CNN for integrated spatio-temporal processing, which are deployed for specific data types and recognition challenges.

These patterns suggest that model selection in the reviewed literature increasingly depends on combining complementary capabilities rather than on choosing a single architectural family, a shift that the multidimensional evaluation in the following subsection makes explicit.

4.3. Comparative Performance Analysis and Architectural Trade-Offs

This subsection uses the established taxonomy to provide a multidimensional and comparative evaluation of the nine architectural families. The four dimensions considered are raw classification accuracy, computational efficiency for edge deployment, robustness to environmental variations, and effectiveness in fine-grained activity recognition, because real-world HAR deployment requires balancing several priorities rather than optimizing a single metric.

These axes were selected from recurring evaluation criteria in the reviewed studies and from established HAR deployment concerns in the literature. Because cross-study accuracy comparison is methodologically inconsistent across heterogeneous datasets, the framework assigns qualitative ratings through cross-corpus synthesis rather than through a quantitative leaderboard.

Table 3 summarizes the evidence from the reviewed literature into qualitative ratings to make these trade-offs explicit. The labels High, Medium, and Low indicate comparative

tendencies across the reviewed studies rather than absolute rankings. The “Supporting Evidence” column summarizes the main interpretation associated with each family.

Table 3. Summary of Multi-Dimensional Performance Evaluation Across Architectural Families

Family	Performance	Edge Devices	Environmental Generalization	Fine-Grained Activity	Supporting Evidence
Classical ML models	Medium	High	Low	Low	Effective in constrained settings and easy to deploy, but performance depends heavily on handcrafted features and re-engineering for new scenes [14,33,37,46]. Strong spatial encoders and common backbones for HAR pipelines, but temporal reasoning often has to be added by another module [12,32,34,40,42]. Joint spatio-temporal learning improves motion capture and clip-level understanding, although this comes at substantial computational cost [25,31,43,47]. Temporal memory helps distinguish activities with similar frames but different orderings, while sequential execution limits speed [20,28,43,47]. Parallel temporal convolutions offer efficient long-range sequence modelling and are attractive for resource-constrained deployment [21,34,38]. Self-attention improves global context modelling and selective focus, but parameter cost and complexity remain deployment challenges [2,22,23,35,42]. Skeleton priors improve structural reasoning, robustness, and privacy preservation, although success depends on reliable pose estimation [9,23,26]. Best suited to multi-actor or group activities, where explicit modelling of interpersonal relations is more important than generic scene classification [9,41]. Combining complementary modules gives the strongest overall trade-off in the literature, at the cost of higher design and optimization complexity [1,2,7,9,20].
CNN architectures	High	High	Medium	Medium	
3D CNN models	High	Low	Medium	High	
Recurrent architectures	Medium	Medium	Medium	High	
TCN models	Medium	High	Medium	High	
Attention-Based models	High	Low	High	High	
Graph-Based models	High	Medium	High	High	
Relational models	Medium	Medium	Medium	High	
Hybrid architectures	High	Medium	High	High	

Note: The qualitative ratings synthesize comparative tendencies reported in the reviewed literature and should not be interpreted as universal rankings independent of dataset, modality, or deployment setting.

Building on the synthesis in Table 3, the following analysis identifies the main trade-offs and determines which architectural families are best suitable to different performance priorities.

Regarding classification accuracy on benchmark datasets, three architectural categories demonstrate prominent results. Hybrid architectures exhibit high performance levels, as the integration of different models facilitates the capture of spatio-temporal dynamics [2,7,12]. For instance, tri-models such as DenMobVit have established performance benchmarks, surpassing existing models with accuracy increases of up to 2.9% on standard datasets [2]. Attention-based models, specifically ViT, demonstrate capacity for tasks involving long-range dependencies within a sequence [2,22,23]. These models achieve precision rates where base ViT configurations reach accuracies of 95.07% [2]. Furthermore, graph-based models are utilised for skeleton-based recognition, as they utilise the structural relationships between body joints [2,25]. Architectures such as ST-GCN report training accuracies of up to 98.9% on pose data [2,25].

Requirements for edge deployment on resource-constrained devices differ significantly from standard implementations. In these contexts, architectures developed for computational efficiency are prioritised. CNN variants, specifically lightweight models such as MobileNet and EfficientNet, are frequently selected due to the trade-off between execution speed and accuracy [12,21,34,42]. MobileNetV3 variants are utilised in resource-limited scenarios, achieving top-1 accuracies of 91.5% with storage requirements ranging from 35 to 350 MB [2]. TCN serve as an alternative to recurrent models, as their parallelised architecture facilitates low inference latency for temporal modelling [21,34]. Although not consistently lightweight, graph-based models offer a potential approach via distributed computing, where skeleton extraction is performed at the edge while computationally intensive processing is managed in the cloud [2].

Environmental generalisation across diverse environments is necessary for robustness in practical applications. Hybrid architectures demonstrate effectiveness in this regard, as integrated feature learning results in representations that generalise across datasets [2,7]. Attention-based models also exhibit generalisation capabilities; specifically, self-attention mechanisms allow these models to adapt to varying contexts by emphasizing salient features [35]. Similarly, graph-based models display cross-dataset robustness, as skeleton-based inputs are invariant to changes in appearance and background [2,41].

Activity recognition between similar classes (for example, “drinking water” versus “brushing teeth”) requires the identification of motion nuances [2,7]. Hybrid architectures are applied to this challenge, as they combine spatial, temporal, and contextual features to capture movement patterns. Attention-based models are also utilised in this context, as dynamic feature weighting allows for a focus on discriminative body parts or temporal segments [2,35]. Furthermore, relational and graph-based models are employed for social interactions and joint movement analysis, where modelling the relationships between entities is fundamental [6,41]. Frameworks such as GCN-LSTM-Attention have achieved prediction accuracies of 94.32% by capturing the relational dynamics between multiple actors [9].

These persistent challenges, ranging from environmental generalization to the high computational costs of complex models, require a critical evaluation of how different architectural strategies prioritize competing requirements.

The comparison shows a hierarchical specialization framework that breaks the performance landscape into three strategic tiers:

1. Hybrid architectures demonstrate high performance levels in applications where computational cost is not a primary constraint. These models integrate spatial and temporal components to achieve high accuracy and robustness.
2. Optimised CNN and TCN architectures provide a trade-off between efficiency and accuracy for real-time applications operating under resource constraints.
3. Specialised models are utilised for specific technical requirements. Attention-based models are applied to long-range temporal dependencies, while graph-based models are used for skeleton-based recognition.

This analysis shows that the field has matured beyond seeking a universally superior architecture. The selection of a model has changed into selecting the one that matches performance with the application setting and the goal to be achieved when employing it into development.

This shift is driven by an evolution in thinking, as the focus on one better architecture has shifted to aligning models’ strengths with application constraints. The trade-off space now includes trustworthiness and explainability, in addition to accuracy and efficiency.

Hybrid and attention-based models report among the highest accuracies in the reviewed studies through their multi-stage design and self-attention mechanisms, generating

intricate, high-capacity decision boundaries. However, their complexity presents technical and regulatory challenges in fields where decisions must be auditable and justifiable.

Thus, architectural choices are influenced by value judgments about predictive power, transparency, and the right to an explanation, re-framing model selection as a social and technical decision.

4.4. Contextual Mapping and Application Domains

After the comparative analysis it is important to map the technological developments into different application domains of HAR, where the goal is to understand how domain-specific requirements influence architectural innovation. Accordingly, the existing research can be broadly classified into six distinct application categories:

- **Education:** Systems are developed for learning environments, such as recognizing student activities in classrooms or during online education.
- **Healthcare:** Systems are designed for patient care, including elderly monitoring, rehabilitation exercise tracking, and clinical assessment of medical conditions.
- **Industrial:** Applications aimed at monitoring workers' activities to ensure safety protocols are followed and to prevent accidents in settings such as construction sites or factories.
- **Security:** Applications focused on safety and threat detection, such as violence recognition, suspicious behaviour identification, and real-time monitoring in public or private spaces.
- **Sports:** Systems are designed to recognize, classify, and analyzes athletic activities and movements for training or performance assessment.
- **Generic:** Research is focused primarily on methodological advancements and benchmark performance, without a specific, primary end-use case driving the article.

Figure 6 categorises the application domains identified across the reviewed literature.

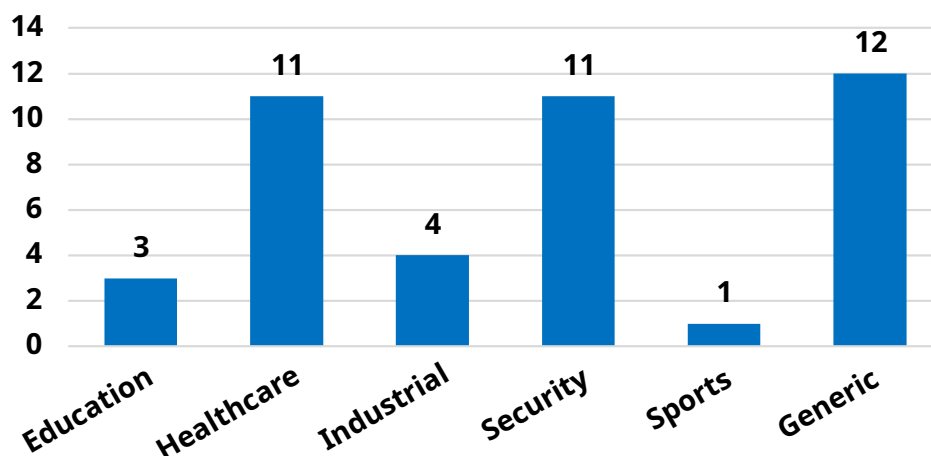


Figure 6. Distribution of Application Contexts.

Analysis of these findings provides the foundation for the subsequent discussion regarding the three most frequent technological implementations. The *Healthcare domain* represents a major field for HAR, since there is an increasing focus on automated health assistance and remote patient monitoring [2,10].

A closer examination of the 11 healthcare-focused articles reveals distinct priorities within this domain. As seen in Table 1, the majority, 6 articles, are dedicated to “People Monitoring”, focusing on the supervision of elderly individuals or people with disabilities in homes, nursing homes, or hospitals. This is followed by 3 articles on “Condition

Monitoring”, where the goal is to detect or assess specific medical conditions, such as Parkinson’s disease. The remaining 2 articles focus on “Physical Procedures”, which include applications like physiotherapy exercise tracking or assisting in surgical procedures.

This distribution suggests that the primary research effort is directed towards continuous, non-intrusive monitoring of individuals in daily living situations. The need for such systems to operate reliably in complex, real-world home environments poses significant challenges in model robustness and privacy preservation, influencing architectural choices toward privacy-aware models like those based on skeleton data [2,10,37].

Interestingly, the smaller number of articles focused on “Physical Procedures” (such as rehabilitation exercises) does not indicate a lack of importance, but rather highlights a different set of technical challenges. These applications often require fine-grained activity recognition to distinguish between correct and incorrect exercise execution, demanding higher spatial and temporal precision from models. The adoption of high-accuracy models in these clinical or therapeutic settings is particularly hindered by their “black-box” nature [2,27]. Clinicians and patients need not only correct predictions but also justifications for them [27,44].

In this context, graph-based and skeleton-based models offer a compelling trade-off, as they operate on abstract joint coordinates rather than raw RGB pixels, preserving visual privacy [2,44], and additionally, they are more robust than RGB-based models because they are invariant to lighting, clothing, and background clutter [2,25]. These advantages are particularly valuable for “People Monitoring” in uncontrolled home environments. The emergence of “Condition Monitoring” research further validates the use of skeleton data, as subtle changes in gait or posture, captured through joint trajectories, can serve as digital biomarkers for conditions like Parkinson’s disease.

The Security domain is a major source of new ideas, especially for applications requiring real-time analysis and proactive threat detection [12]. Research in this area is driven by the need to improve public safety and security in different settings [2,10,12]. These applications require that the models be reliable, that the speeds of making decisions be high, and that they work well in environments that are always changing, with different lighting, hidden parts, and viewpoints [12,29,43].

A significant segment of the articles is categorized as Generic domain, where the focus is on the methods used rather than on how they are applied in specific situations [21,26]. These are very important for advancing the field because they focus on new architectural innovations, improving the performance of models on standard datasets, and achieving new results that are among the best [2,12,21,43]. While it doesn’t focus on one specific use, the progress made in this area provides the essential tools and models that are later adapted and refined for the specialized fields of healthcare and security [7,12,43].

The research shows that HAR research is not spread out evenly, but is instead focused on areas that are important to society. Two main reasons explain this concentration:

- Pressing global issues, such as ageing populations, rising healthcare costs, and the demand for enhanced public safety, have created an urgent need for automated monitoring solutions that can be easily scaled up or down.
- Strict requirements of the domains, like the need for strength, reliability, and privacy, encourage new innovations. They work on making better models that can be used in more situations and that we can trust.

As a result, the HAR research landscape is not defined by a uniform spread of applications. Instead, it is defined by a targeted concentration on areas where the technology can deliver significant societal and economic value. This cements its role as a critical enabling technology.

5. Future Research Directions

The proposed future directions follow directly from the evidence summarized in Table 3 and the application patterns discussed in Section 4. The analysis shows that Hybrid architectures currently provide the strongest overall balance of accuracy, robustness, and fine-grained recognition [1,2,7,9,20], while Attention-based models also achieve strong results when long-range dependencies are important [2,22,23,35,42]. However, the same analysis shows that these families remain less suitable for edge deployment than lightweight CNN and TCN pipelines, creating a clear high-accuracy versus low-efficiency conflict. Therefore, future research should focus on lightweight fusion strategies, efficient Transformer variants, model compression, knowledge distillation, and hardware-aware neural architecture search. This will allow the performance gains of systems such as DenMobVit to be retained without sacrificing deployability [2,21,34,38].

A second priority arises from the gap between benchmark success and operational robustness. Section 4 showed that hybrid models generalize better than isolated architectures [2,7], and that graph-based or skeleton-based approaches are especially attractive in privacy-sensitive and cluttered environments because they are less affected by lighting and background variation [2,25,41,44]. At the same time, graph-based systems still depend on reliable pose extraction, and security-oriented models must operate under low light, occlusion, and viewpoint variation [2,12,25,29,43]. Therefore, future research should prioritize larger and more diverse datasets, domain randomization, unsupervised domain adaptation, and evaluation protocols that better reflect real deployment conditions instead of controlled laboratory settings.

The domain analysis also indicates that future research should be guided by application-specific constraints rather than by the search for a universally superior model. In healthcare, most of the reviewed work is concentrated in people monitoring and condition monitoring, where privacy preservation and robustness have favoured skeleton-based and graph-based representations [2,10,25,37,44]. By contrast, rehabilitation and clinical assessment require fine-grained discrimination, yet the analysis also showed that the most accurate families often behave as black boxes, limiting trust in clinical settings [2,27,44]. This suggests two related areas of future research: privacy-aware longitudinal monitoring systems for real-world care environments, and interpretable high-precision models for rehabilitation and condition assessment.

More broadly, the review suggests that progress in HAR will depend on resolving explicit trade-offs rather than proposing increasingly complex architectures in isolation. The contrast between high-performing hybrid, attention-based, and 3D CNN families [2,31,42,43,47] and the more deployment-friendly CNN and TCN families [21,34,38] shows that efficiency remains a central design constraint, while the success of relational and graph-based hybrids in group activity recognition [9,41] shows that richer reasoning often comes at the cost of greater system complexity. Future research should therefore explore adaptive architectures that can adjust complexity to context, integrate explainability from the design stage, and validate models through closer collaboration between academic teams and deployment-oriented partners such as SafeCaring [8].

6. Conclusions

This literature review comprised three core objectives: cataloguing predominant computer vision models for HAR, critically evaluating their performance trade-offs, and identifying their primary application contexts. This was achieved through a PRISMA-guided analysis of 42 articles published between 2021 and 2025.

The main contributions are a reproducible taxonomy of nine architectural families, a multidimensional benchmark comparison based on accuracy, efficiency, environmental

generalisation and fine-grained discrimination, and a mapping of technology choices to domain-specific needs. The transparent, evidence-based methodological approach offers practical guidance for model selection in real-world deployment.

Nevertheless, the findings of this review are subject to limitations that should be acknowledged. The search strategy was limited to four major databases and a publication window from January 2021 to December 2025, potentially excluding relevant work from other sources. Additionally, the review is descriptive rather than prescriptive, meaning that it identifies trends, trade-offs and gaps in the existing literature, rather than proposing a novel model or empirically validating any architectural claims. Given the rapid pace of innovation in deep learning, particularly in areas such as vision transformers and efficient attention mechanisms, the findings represent a snapshot of a dynamic and fast-moving field.

In conclusion, this review confirms that HAR has evolved into an application-driven discipline, in which the selection of models is not about identifying a universally superior architecture, but rather about aligning the strengths of architectures with specific deployment constraints and societal challenges. The transition from academic prototypes to real-world systems, as exemplified by industry solutions such as SafeCaring, will also require sustained efforts in terms of robustness, explainability and computational efficiency.

Author Contributions: Conceptualization, L.H.T., E.R. and F.B.P.; Methodology, L.H.T., E.R., J.B. and F.B.P.; Investigation, L.H.T. and E.R.; Writing—original draft preparation, L.H.T. and E.R.; Writing—review and editing, L.H.T., E.R., J.B. and F.B.P.; Supervision, E.R., J.B. and F.B.P.; Funding acquisition, E.R. and F.B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326—Centre for Informatics and Systems of the University of Coimbra. This research was co-funded by the European Regional Development Fund (ERDF), through the Centro2030 Programme, under project no. 17358, “Safecaring—Enhancing care safety and quality through reduced administrative burden”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors acknowledge SafeCaring for providing practical insights into real-world healthcare challenges that motivated aspects of this research. During the preparation of this work, the authors used AI-assisted language tools to improve readability and grammar. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Author Edite Ravella was employed by the company SafeCaring. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Arasi, M.A.; Alkudhayr, H.; Alkharashi, A.; Alneil, A.A. Enhancing indoor activity recognition for disabled persons using multi head self attention recurrent neural network with improved pelican algorithm. *Sci. Rep.* **2025**, *15*, 33164. [[CrossRef](#)] [[PubMed](#)]
2. Zaher, M.; Ghoneim, A.S.; Abdelhamid, L.; Atia, A. Fusing CNNs and attention-mechanisms to improve real-time indoor Human Activity Recognition for classifying home-based physical rehabilitation exercises. *Comput. Biol. Med.* **2025**, *184*, 109399. [[CrossRef](#)] [[PubMed](#)]
3. Hirooka, K.; Hasan, M.A.M.; Shin, J.; Srizon, A.Y. Ensembled Transfer Learning Based Multichannel Attention Networks for Human Activity Recognition in Still Images. *IEEE Access* **2022**, *10*, 47051–47062. [[CrossRef](#)]
4. Berger, J.; Lu, S. A Multi-camera System for Human Detection and Activity Recognition. In *Procedia CIRP*; Elsevier B.V.: Amsterdam, The Netherlands, 2022; Volume 112, pp. 191–196. [[CrossRef](#)]

5. de Lope, J.; Graña, M. Deep transfer learning-based gaze tracking for behavioral activity recognition. *Neurocomputing* **2022**, *500*, 518–527. [\[CrossRef\]](#)
6. Tsai, C.Y.; Khosiin, M.W.; Lin, J.J.; Chen, C.S. Multi-granular crew activity recognition for construction monitoring. *Autom. Constr.* **2025**, *179*, 106428. [\[CrossRef\]](#)
7. Ahmed, M.F.; He, G.; Wang, S. TriFusion hybrid model for human activity recognition. *Signal Image Video Process.* **2024**, *18*, 8469–8476. [\[CrossRef\]](#)
8. SafeCaring. The Most Powerful Solution for Healthcare Teams. AI-Based Sensor System for Automated Clinical Documentation, Fall Prevention, and Care Reminders. 2026. Available online: <https://safecaring.pt/> (accessed on 17 June 2026).
9. Tejonidhi, M.R.; Aravinda, C.V.; Kumar, S.V.; Madhu, C.K.; Vinod, A.M. Optimizing Group Activity Recognition with Actor Relation Graphs and GCN-LSTM Architectures. *IEEE Access* **2025**, *13*, 55957–55969. [\[CrossRef\]](#)
10. Sun, T.; Lian, C.; Dong, F.; Shao, J.; Zhang, X.; Xiao, Q.; Ju, Z.; Zhao, Y. Skeletal joint image-based multi-channel fusion network for human activity recognition. *Knowl.-Based Syst.* **2025**, *315*, 113232. [\[CrossRef\]](#)
11. Zhang, C.; Cao, K.; Lu, L.; Deng, T. A multi-scale feature extraction fusion model for human activity recognition. *Sci. Rep.* **2022**, *12*, 20620. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Hussain, A.; Khan, S.U.; Khan, N.; Shabaz, M.; Baik, S.W. AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107218. [\[CrossRef\]](#)
13. Nguyen, P.D.; Le, N.T.; Bui, K.H.; Nguyen, H.Q.; Nguyen, H.Q.; Le, T.L. A method for continuous student activity recognition from classroom videos. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 7913–7937. [\[CrossRef\]](#)
14. Sarveshwaran, V.; Joseph, I.T.; Maravarman, M.; Karthikeyan, P. Investigation on Human Activity Recognition using Deep Learning. In *Procedia Computer Science*; Elsevier B.V.: Amsterdam, The Netherlands, 2022; Volume 204, pp. 73–80. [\[CrossRef\]](#)
15. Raj, R.; Kos, A. An improved human activity recognition technique based on convolutional neural network. *Sci. Rep.* **2023**, *13*, 22581. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Trelinski, J.; Kwolek, B. CNN-based and DTW features for human activity recognition on depth maps. *Neural Comput. Appl.* **2021**, *33*, 14551–14563. [\[CrossRef\]](#)
17. Gawande, U.; Hajari, K.; Golhar, Y. Novel person detection and suspicious activity recognition using enhanced YOLOv5 and motion feature map. *Artif. Intell. Rev.* **2024**, *57*, 16. [\[CrossRef\]](#)
18. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015; pp. 4489–4497. [\[CrossRef\]](#)
19. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of the Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
20. Domingo, J.D.; Gómez-García-Bermejo, J.; Zalama, E. Improving Human Activity Recognition Integrating LSTM with Different Data Sources: Features, Object Detection and Skeleton Tracking. *IEEE Access* **2022**, *10*, 68213–68230. [\[CrossRef\]](#)
21. Shin, J.; Hasan, M.A.M.; Maniruzzaman, M.; Nishimura, S.; Alfarhood, S. Video-Based Human Activity Recognition Using Hybrid Deep Learning Model. *CMES-Comput. Model. Eng. Sci.* **2025**, *143*, 3615–3638. [\[CrossRef\]](#)
22. Wensel, J.; Ullah, H.; Munir, A. ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos. *IEEE Access* **2023**, *11*, 72227–72249. [\[CrossRef\]](#)
23. Han, H.; Zeng, H.; Kuang, L.; Han, X.; Xue, H. A human activity recognition method based on Vision Transformer. *Sci. Rep.* **2024**, *14*, 15310. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2–7 February 2018; Volume 32, pp. 7444–7452. [\[CrossRef\]](#)
25. Tsinikos, V.; Pastaltzidis, I.; Karakostas, I.; Dimitriou, N.; Valakou, K.; Margetis, G.; Stefanidis, C.; Tzovaras, D. Real-Time Activity Recognition for Surveillance Applications on Edge Devices. In *ACM International Conference Proceeding Series*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 293–299. [\[CrossRef\]](#)
26. Sarapata, G.; Dushin, Y.; Morinan, G.; Ong, J.; Budhdeo, S.; Kainz, B.; O’Keeffe, J. Video-Based Activity Recognition for Automated Motor Assessment of Parkinson’s Disease. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5032–5041. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Dutt, M.; Goodwin, M.; Omlin, C.W. An Interpretable Deep Learning-Based Feature Reduction in Video-Based Human Activity Recognition. *IEEE Access* **2024**, *12*, 187947–187963. [\[CrossRef\]](#)
28. Pandey, A.; Kumar, P. BGRU-MTRA: Bilinear GRU networks with multi-path temporal residual attention for suspicious activity recognition. *Neural Comput. Appl.* **2025**, *37*, 185–212. [\[CrossRef\]](#)
29. Saba, T.; Rehman, A.; Latif, R.; Fati, S.M.; Raza, M.; Sharif, M. Suspicious Activity Recognition Using Proposed Deep L4-Branched-Actionnet with Entropy Coded Ant Colony System Optimization. *IEEE Access* **2021**, *9*, 89181–89197. [\[CrossRef\]](#)

30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Yuvaraj, R.; Fogarty, J.S.; Rajalakshmi, R.; Sarkar, R. Classroom activity recognition using hybrid 3D-CNNs and visualization of action features with Grad-CAM. *Neurocomputing* **2025**, *657*, 131694. [\[CrossRef\]](#)
32. Rani, M.; Kumar, M. Efficient Human Activity Recognition Using a Hybrid MobileNetV2-CNN Model. *Natl. Acad. Sci. Lett.* **2025**, *48*, 1720. [\[CrossRef\]](#)
33. Pandurangan, S.; Papandrea, M.; Gelsomini, M. Fine-Grained Human Activity Recognition—A new paradigm. In *iWOAR '22: Proceedings of the 7th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*; Association for Computing Machinery (ACM): New York, NY, USA, 2023; pp. 1–8. [\[CrossRef\]](#)
34. Wei, P.; Ahmedt-Aristizabal, D.; Gammulle, H.; Denman, S.; Armin, M.A. Vision-based activity recognition in children with autism-related behaviors. *Heliyon* **2023**, *9*, e16763. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Alzahrani, A.; Alghamdi, A.M. A vision transformer with recurrent neural network-based fall activity recognition system for disabled persons in smart IoT environments. *Sci. Rep.* **2025**, *15*, 33187. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Alohal, M.A.; Alzahrani, M.Y.; Alghamdi, A.M.; Yaseen, I. Advanced smart human activity recognition system for disabled people using artificial intelligence with snake optimizer techniques. *Sci. Rep.* **2025**, *15*, 31372. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Gaikwad, S.; Bhatlawande, S.; Shilaskar, S.; Solanke, A. A computer vision-approach for activity recognition and residential monitoring of elderly people. *Med. Nov. Technol. Devices* **2023**, *20*, 100272. [\[CrossRef\]](#)
38. Ramesh, S.; Dall'Alba, D.; Gonzalez, C.; Yu, T.; Mascagni, P.; Mutter, D.; Marescaux, J.; Fiorini, P.; Padoy, N. TRandAugment: Temporal random augmentation strategy for surgical activity recognition from videos. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 1665–1672. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Vyas, S.; Vyas, C.; Sharotry, A.; Jimenez, J.A.; Qasem, A.; Mediavilla, F.A.M. Human Activity Recognition in MMH: Pilot Study on Lifting and Lowering Classification. *IFAC-PapersOnLine* **2025**, *59*, 763–768. [\[CrossRef\]](#)
40. Obaid, S.; Bawany, N.Z.; Tahzeeb, S.; Qamar, T.; Mughal, M.H. Deep Learning-Based Driver Activity Recognition Using Diverse Driver Profiles. *IEEE Access* **2024**, *12*, 187192–187209. [\[CrossRef\]](#)
41. Wang, C.; Mohamed, A.S.A. Attention Relational Network for Skeleton-Based Group Activity Recognition. *IEEE Access* **2023**, *11*, 129230–129239. [\[CrossRef\]](#)
42. Chatterjee, R.; Choudhury, R.R.; Gourisaria, M.K.; Banerjee, S.; Dey, S.; Sahni, M.; Leon-Castro, E. Temporal-Aware Transformer Approach for Violence Activity Recognition. *IEEE Access* **2025**, *13*, 70779–70790. [\[CrossRef\]](#)
43. Hussain, A.; Khan, S.U.; Khan, N.; Ullah, W.; Alkhayyat, A.; Alharbi, M.; Baik, S.W. Shots segmentation-based optimized dual-stream framework for robust human activity recognition in surveillance video. *Alex. Eng. J.* **2024**, *91*, 632–647. [\[CrossRef\]](#)
44. Yousuf, M.J.; Lee, B.; Asghar, M.N.; Ansari, M.S.; Kanwal, N. Unlocking Trust: Advancing Activity Recognition in Video Imagery. *IEEE Access* **2024**, *12*, 176799–176817. [\[CrossRef\]](#)
45. Gawande, U.; Hajari, K.; Golhar, Y. Real-Time Deep Learning Approach for Pedestrian Detection and Suspicious Activity Recognition. In *Procedia Computer Science*; Elsevier B.V.: Amsterdam, The Netherlands, 2023; Volume 218, pp. 2438–2447. [\[CrossRef\]](#)
46. Monika; Bhatia, M. Extraction of posture silhouettes using human posture feature points and spatial masks for activity recognition in public place. *J. Eng. Res.* **2022**, *10*, 77–88. [\[CrossRef\]](#)
47. Ullah, A.; Muhammad, K.; Ding, W.; Palade, V.; Haq, I.U.; Baik, S.W. Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. *Appl. Soft Comput.* **2021**, *103*, 107102. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.