Vol. 12, No. 2 (2026), pp.10–36

CHITKARA UNIVERSITY

Journal of Multidisciplinary Research in Healthcare



Optimization-Driven Deep Learning Models for 3D Human Action Recognition: A Survey

Kamalpreet Kaur[1]*, Ankit Bansal[1] and Baljit Singh Khehra[2]

[1]*Department of Computer Science & Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.*

[2]*Department of Computer Science & Engineering, Jagat Guru Nanak Dev Punjab State Open University, Punjab, India.*

*kamalpreethind@gmail.com (Corresponding Author)

RESEARCH ARTICLE **Open Access**

ARTICLE INFORMATION

Received: January 28, 2025
Accepted: April 29, 2026
Published Online: July 05, 2026

Keywords:
Human action recognition, 3D skeletal data, Deep learning, Graph convolutional networks, Metaheuristic optimization, Spatiotemporal modeling, Hybrid neural networks, Transformer-based models

ABSTRACT

Background: Due to its numerous applications in healthcare, surveillance, human-computer interaction, sports analytics, and smart environments, Human Action Recognition (HAR) using 3D skeletal data has grown in importance as a field of study. Conventional HAR techniques mainly relied on manually created features, which were unreliable and data dependent. The ability to model intricate spatiotemporal patterns in human motion has greatly improved with the development of deep learning models like CNNs, RNNs, GCNs, and hybrid architectures. The integration of optimization techniques into deep learning-based HAR systems is motivated by unresolved issues like high computational cost, sensitivity to hyperparameters, poor cross-dataset generalization, and lack of interpretability.

Purpose: This survey's goal is to offer a thorough analysis of optimization-driven deep learning models for 3D human action recognition. The research seeks to examine cutting-edge deep learning architectures for 3D HAR and analyze how metaheuristic optimization techniques can enhance the precision, effectiveness, and generalization of models.

Methods: Using a methodical survey approach, this study covers deep learning architectures such as Transformer-based, CNN-based, RNN-based, GCN-based, and Hybrid-DNN models. Optimization techniques include Chaos Game Optimization, Whale Optimization Algorithm, Grey Wolf Optimizer, Rao-3, Wild Horse Optimization, and Sea Horse Optimization. Model comparison is conducted using benchmark datasets such as NTU RGB+D, NTU RGB+D 120, Kinetics-Skeleton, SYSU-3D, and N-UCLA. Assessment uses standard performance metrics, mainly accuracy, along with reported robustness and computational efficiency.

Results: According to the survey, optimization-driven deep learning models routinely perform better in 3D HAR than conventional and non-optimized methods. Higher recognition accuracy is attained by optimized models, frequently surpassing 90–95% on benchmark datasets. Hyperparameter tuning is greatly enhanced by metaheuristic optimization, which lowers overfitting and computational inefficiency. Despite advancements, problems like cross-dataset generalization, model explainability, and real-time processing still exist.

Conclusion: Recent developments in optimization-driven deep learning models for 3D Human Action Recognition (HAR) were examined in this survey. The analysis demonstrates that combining metaheuristic optimization techniques with deep learning architectures greatly increases recognition accuracy, generalization, and computational efficiency.



DOI: 10.15415/jmrh.2026.122002

1. Introduction

The field of Human Activity Recognition (HAR) deals with the detection and interpretation of human behaviours from visual data sources, mainly images or video frames (Sundaresan *et al.*, 2025). Applications in a variety of domains, such as healthcare, smart environments, sports analytics, surveillance and security systems, autonomous driving systems, human–computer interaction (HCI), and human–robot interaction (HRI) (Zhou *et al.*, 2020; He *et al.*, 2020; Bian *et al.*, 2024), are made possible by the ability to autonomously identify activities from unprocessed visual input. HAR enables remote patient monitoring in the healthcare industry (Hoang *et al.*, 2023), and it may

©Author(s) 2026. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are also included in the article's Creative Commons licence. To view a copy of this licence, visit https://creativecommons.org/licenses/by/4.0/

optimise energy use in smart environments (Tien *et al.*, 2021). HAR improves sports performance analysis in sports analytics (Zhang *et al.*, 2017) and helps identify suspicious behaviour in security monitoring (Yu *et al.*, 2016). These varied applications demonstrate the expanding influence of HAR across a wide range of domains.

Based on the type of input data and the equipment required to collect data on human behaviours, HAR may be broadly divided into two categories (Figure 1). They are:

- **Sensor-Based:** To collect useful input on human motion and activity in the form of signals, sensors are either directly or indirectly attached to the subject's body or placed in the surrounding environment. One excellent example of using wearable, non-visual sensors to collect HAR data is inertial sensor-based HAR (Zahin *et al.*, 2019).
- **Vision-Based:** RGB cameras and other visual sensors are commonly used to capture visual data in the form of 2D images (Zahin *et al.*, 2019) or 3D data (RGB-D sensors). After some pre-processing to eliminate ambiguities, the feature-rich visual data can be used to train the classification model.

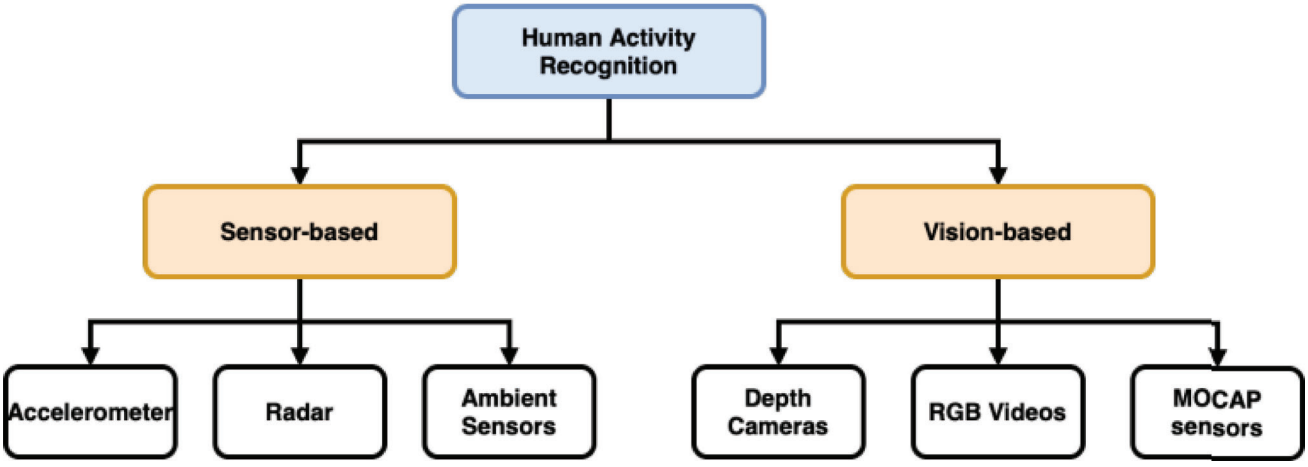


Figure 1: A General Hierarchy of HAR Methods

Depending on the sensor used to gather the data, there are basically two kinds of HAR systems: sensor-based HAR and vision-based HAR. However, this research's primary emphasis is on HAR utilising 3D data and the developments in this area of study in the contemporary period. The classification problem known as 3D HAR involves labelling a video according to the action that an actor or a group of actors is performing. Due to a number of fundamental issues with data collection, feature extraction, and model selection for classification, 3D HAR is one of the most complex and demanding problems in computer science (Meng *et al.*, 2020). Due to a number of initial difficulties with the video stream, including target distance, poor illumination, unstable video, low resolution, flickering video, and the high cost of equipment and sensors, visual sensor-based 3D HAR, in particular, has some limitations.

Recent developments in deep learning (DL) and machine learning (ML) technologies have greatly improved the accuracy and scalability of HAR systems. Machine learning techniques were used in many earlier studies to recognise human activities (Lara & Labrador, 2013). They heavily depend on statistical methods (Bulling *et al.*, 2014), symbolic representation (Lin *et al.*, 2003), and time-frequency transformation as feature extraction strategies. Nonetheless, the retrieved features are heuristic and meticulously designed. There were no universal or systematic feature extraction techniques to efficiently collect distinctive properties for human activities.

DL has shown notable success in recent years in modelling complex abstractions from complex data (Pouyanfar *et al.*, 2018) in a variety of fields, including voice processing, computer vision, and natural language processing. Related research emerged in this field after early publications, such as (Ha *et al.*, 2015), investigated the efficacy of deep learning in human activity identification. Recent efforts are being made to solve these particular obstacles, in addition to the inevitable advancement of deep learning in human action identification.

Since DL carries out end-to-end optimisation, it is particularly useful for classification issues in which related activities benefit from one another (transfer learning). In order to achieve impressive results in both 2D and 3D HAR

challenges, a number of contemporary action recognition techniques (Guha *et al.*, 2021) use deep Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Convolutional Networks (GCNs) (Ye *et al.*, 2020). However, because of its sudden success, flurry of innovation, and lack of theoretical backing, deep learning is still met with reluctance by academics. Since knowledge of a particular traversal of the joints is necessary to determine the spatial structure of the skeleton in 3D HAR, a disadvantage of utilising RNNs is that they are unable to capture all spatial relationships. GCNs may not be sufficient for sophisticated 3D HAR tasks because of scalability concerns associated with their number of nodes.

Researchers are increasingly using optimisation approaches to overcome these restrictions and improve the performance of DL models (Basak *et al.*, 2022; Ozcan & Basturk, 2020; Usmani *et al.*, 2021). The choice of hyperparameters has a direct influence on model accuracy and generalisation, making classification algorithms very sensitive to their performance. However, because of the high dimensionality and complexity of the search space, finding the ideal collection of hyperparameters is still a challenging process. Because they provide effective methods for hyperparameter tuning in deep learning-based HAR models and increase their overall success rate, metaheuristic optimisation algorithms have become essential tools for automating this process.

Human motion analysis, motion analysis in video data, posture prediction, HAR from 3D data, depth data, vision-based data, sensor-based HAR, wearable sensor-based HAR, human pose estimation, skeleton-based HAR, and RFID-based HAR are just a few of the HAR aspects that have been the subject of countless reviews over the last ten years (Basak *et al.*, 2022; Ozcan & Basturk, 2020; Usmani *et al.*, 2021). Furthermore, the various datasets and modalities employed in HAR have been examined in a number of papers (Singh *et al.*, 2020; Pareek & Thakkar, 2021).

Deep learning methods in 3D Human Action Recognition (HAR) have been reviewed in a number of recent survey studies, but each has significant drawbacks. The study by Karthickkumar and Kumar (2020) mentions DL-based HAR in passing, including a helpful dataset table, but falls short in terms of in-depth discussion of the HAR pipeline and classification methods. Although Ren *et al.* (2024) provide a thorough analysis of skeleton-sequence-based HAR, with a focus on Graph Convolutional Network (GCN) techniques, they neglect to examine application areas or future research prospects and leave out the early stages of the HAR pipeline. The same is true for Rajendran *et al.* (2020), who concentrate on RGB-D data and pre-processing procedures while offering a broad summary of ML and DL approaches. However, they do not address more recent GCN techniques or other data types, such as skeletal or inertial data. Although, Jaiswal and Jaiswal (2020) provide insightful information on new skeleton-based HAR techniques, they ignore pre-classification procedures, other data types, including RGB-D and inertial inputs, and possible applications. Last but not least, Al-Faris *et al.* (2020) provide a complete assessment of feature extraction and classification methods for RGB-D and skeletal data, although they overlook developments in GCN-based HAR and inertial sensor-based modalities.

In the studies by Xing and Zhu (2021); Arshad *et al.* (2022), the authors conducted a full survey of HAR based on the 3D human skeleton, in which the research was surveyed and divided into three approaches: RNN-based, CNN-based, and GCN-based. In these two survey studies, the challenges of HAR based on the 3D human skeleton were not presented and analysed. Islam *et al.* (2022) only conducted the survey using CNN-based HAR. When taken as a whole, these limitations highlight the necessity for a more thorough assessment that addresses the whole HAR pipeline, various data modalities, current developments in deep learning, and real-world applications.

This study addresses the limitations identified in prior surveys, including insufficient pipeline coverage, restricted modality scope, and inadequately examined optimisation strategies, by offering a thorough overview of cutting-edge deep learning methodologies for 3D Human Action Recognition (HAR), accompanied by a detailed investigation of sophisticated metaheuristic optimisation techniques. This dual emphasis tackles significant issues in hyperparameter optimisation, scalability, and model generalisation. The study provides a cohesive and progressive viewpoint by synthesising lessons from various deep learning architectures and optimisation techniques, facilitating the development of resilient, efficient, and scalable HAR systems designed for real-world applications.

2. Literature Review

Human identification based on skeletal data has seen considerable advancement in the last ten years due to technological innovations in sensor technology, deep neural networks, and computing resources. The initial methods were based on hand-crafted features derived from joint angle paths, covariance matrices, and time sequences (Jain *et al.*, 2008; Wang *et al.*, 2013). Although these techniques showed plausible results on limited-size datasets, they lacked the ability to generalize to different environments and types of actions.

The introduction of recurrent neural networks (RNNs) and long short-term memory (LSTM) units marked a shift in sequence-based action recognition. Du *et al.* (2015)

introduced a hierarchical RNN used to learn skeletal joint sequences, and they showed that recurrent architectures could effectively learn temporal dependencies in human motion. Then, Li *et al.* (2019) proposed independently recurrent neural networks (IndRNNs) that can handle long sequences without the gradient vanishing problem. Nevertheless, one of the most fundamental weaknesses of all RNN-based approaches is that these methods simply cannot encode the spatial topology of the human skeleton, as joints are considered discrete elements of the sequence rather than entities that are spatially linked.

Convolutional neural networks were subsequently applied to skeleton-based HAR, where joint coordinates are treated as pseudo-images, and 2D spatial convolutions are used to learn the features of body structures (Caetano *et al.*, 2019). This method was extended to 3D volumetric heatmaps in PoseConv3D (Duan *et al.*, 2022), providing more detailed spatiotemporal representations. Despite their ability to model spatial information, CNN-based approaches also have the disadvantage of information loss during the encoding process, especially with fine-grained motion patterns where inter-joint relational context is important.

Graph Convolutional Networks (GCNs) became the prevailing paradigm of skeleton-based HAR after the groundbreaking study of Yan *et al.* (2018), who proposed ST-GCN, a spatial-temporal graph convolutional network capable of representing the human skeleton as a graph with nodes (joints) and edges (bones) (Yan *et al.*, 2018). Later studies suggested dynamic graph learning (Chen *et al.*, 2021), attention-based GCNs (Ding *et al.*, 2019), and channel-wise topology refinement to overcome the weaknesses of fixed graph structures. Nonetheless, GCN-based models are computationally intensive, and their application in real-time embedded systems remains underexplored.

Hybrid deep learning architectures combining RNNs, CNNs, and GCNs have demonstrated steady improvements in accuracy on benchmark datasets. Multi-stream feature fusion models such as AGC-LSTM (Si *et al.*, 2019), SGN (Zhang *et al.*, 2020), and PSUMNet (Trivedi & Sarvadevabhatla, 2022) show that it is possible to capture local joint-level dynamics and global motion semantics through multi-stream feature fusion. More recently, transformer-based architectures (Cho *et al.*, 2020; Shi *et al.*, 2020) have competed with GCN architectures, using self-attention to capture long-range temporal dependencies, but they require large quantities of training data and consume substantial computational resources.

Recent research has started to incorporate metaheuristic optimization algorithms to automatically tune hyperparameters in deep learning-based HAR systems (Basak *et al.*, 2022; Ozcan & Basturk, 2020; Usmani *et al.*, 2021; Al-Wesabi *et al.*, 2021; Challa *et al.*, 2023; Thakur *et al.*, 2025). Algorithms such as the Whale Optimization Algorithm (WOA), Grey Wolf Optimizer (GWO), and Chaos Game Optimization (CGO) have been used to optimize learning rates, batch sizes, and network depths and have been shown to outperform manually tuned baselines. However, these studies remain inconclusive, as there is no single comparative framework to assess various optimizers under similar experimental conditions for HAR.

Although the current surveys are extensive, some important research gaps still remain, which are summarized in Table 1.

Table 1: Summary of Research Gaps Identified from Prior Survey Studies

Prior Survey / Study Focus	**Identified Research Gap Addressed by This Work**
Prior surveys on RNN/LSTM-based HAR	Temporal modelling only; spatial joint relationships ignored; no optimization coverage
GCN-focused survey studies	Limited to fixed graph topologies; scalability constraints not addressed; real-world deployment gaps
Hybrid DNN reviews	Inadequate analysis of computational cost versus accuracy trade-offs; limited cross-dataset evaluation
Optimization-integrated HAR studies	Focus on individual algorithms; no comparative framework across metaheuristics for HAR-specific tasks
Transformer-based methods	Spatial feature encoding limitations; not evaluated on resource-constrained or real-time HAR scenarios

This survey directly addresses these gaps by:

1) providing a unified review of RNN-, CNN-, GCN-, hybrid-, and transformer-based approaches within a single framework;
2) systematically comparing metaheuristic optimization algorithms applied to 3D HAR;
3) critically evaluating the scalability and real-world applicability of the reviewed methods; and
4) proposing concrete future research directions targeting model explainability, cross-dataset generalization, and lightweight deployment.

Figure 2 presents the systematic process flow underlying this survey.

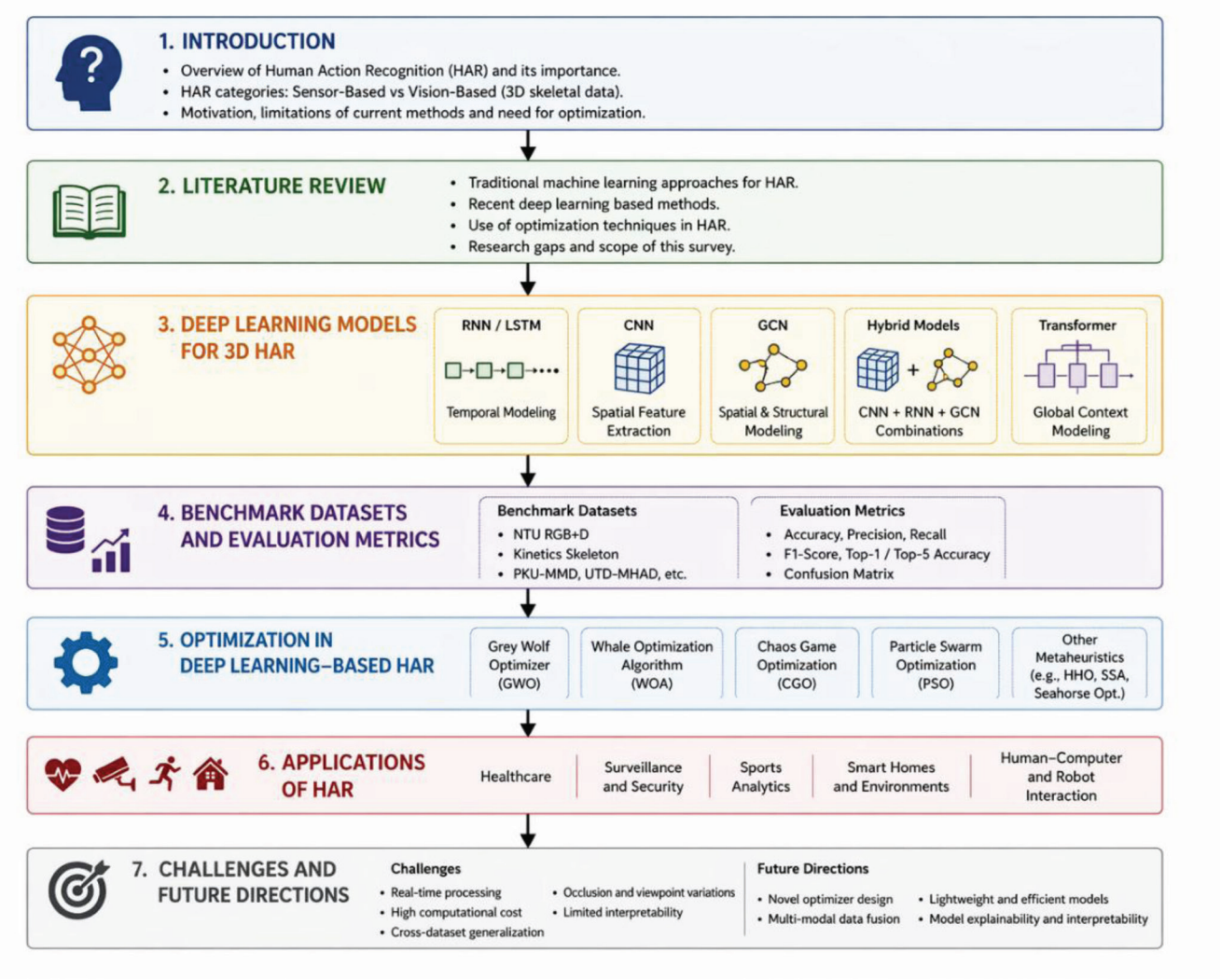


Figure 2: Survey Methodology Process Flow Diagram for the Review of Optimization-Driven Deep Learning Models for 3D Human Action Recognition

3. Deep Learning Architectures for 3D HAR

DL techniques have shown unexpected performance in recent years as they have continued to advance in the majority of current computer vision problems. The deep learning approach based on data-driven features has therefore received a great deal of attention (Figure 3 illustrates the broad structure).

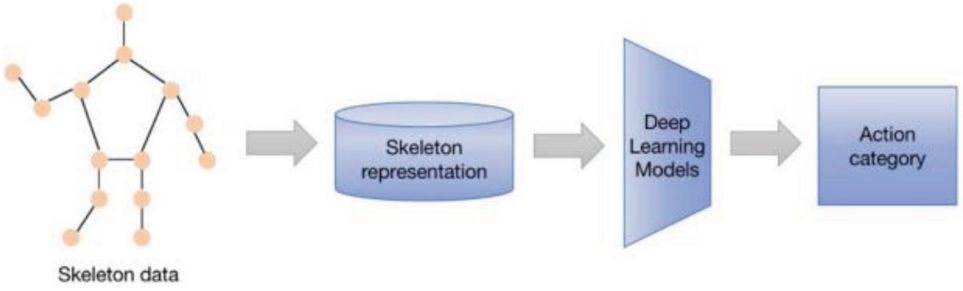


Figure 3: General Framework of Skeleton □ Based Action Recognition Using Deep Learning Methods

In many pattern recognition applications, deep learning architectures may demonstrate remarkable performance by learning hierarchical representations. For example, skeletal data have been modelled for 3D action detection using recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) (Shahroudy *et al.*, 2016; Zhang *et al.*, 2017). Due to their capacity to represent temporal sequences, RNN-based methods perform very well in 3D action recognition tasks; nevertheless, these structures are unable to effectively understand the spatial relationships among the skeleton joints (Yang *et al.*, 2018).

Du *et al.* (2015) suggested a hierarchical framework to capitalise on the spatial linkages. Each skeleton sequence is represented by the authors as a 2D array, where the spatial structure of each frame is represented by rows and the temporal dynamics of the sequence are stored as variations in columns. The representation is then provided to convolutional neural networks (CNNs), which are naturally able to learn structural information from 2D arrays. These representations are very condensed, containing the whole video sequence in a single image. However, the deep relationships between body joints are seldom captured by expressing the skeletal data as a vector series or 2D arrays, leaving out a wealth of valuable motion information.

Some approaches (Li *et al.*, 2019; Shi *et al.*, 2019) build a skeleton graph with vertices representing joints and edges representing bones in order to better acquire joint relationships. They then use graph convolutional networks (GCNs) to extract associated features. Table 2 compares three deep network framework approaches with conventional techniques.

Table 2: Comparison of Traditional Methods and Three Deep Network Framework Methods

Method Type	Summary
Traditional methods (Wang *et al.*, 2012; Zanfir *et al.*, 2013)	Since traditional skeleton-based action detection is considered a challenging task and often requires motion patterns to be extracted from a certain skeleton sequence, handcrafted characteristics have been the subject of in-depth research. Handcrafted characteristics, however, are always data-dependent and superficial.
RNN-based methods (Shahroudy *et al.*, 2016; Zhang *et al.*, 2017)	Skeletal sequences for RNN-based techniques are natural time series of joint coordinate locations that may be thought of as sequence vectors, and the RNN itself can handle time-series data because of its special structure. Even though RNN-based techniques provide excellent results, they are unable to effectively learn the spatial relationships among the skeleton joints.
CNN-based methods (Du *et al.*, 2015; Caetano *et al.*, 2019)	Many studies encode the skeletal joints into multiple 2D pseudo-images in order to specifically investigate spatial information. CNN-based methods have the benefit of naturally learning structural information from 2D arrays (i.e., learning spatial relations from the skeleton joints) thanks to the representations they utilise.
GCN-based methods (Li *et al.*, 2019; Shi *et al.*, 2019)	In a non-Euclidean space, the skeleton is naturally organised like a network, with the joints serving as vertices and the natural connections inside the human body serving as edges. It is challenging to generalise the earlier approaches to skeletons with arbitrary shapes, and they are unable to take advantage of the graph structure of the skeleton data. A series of skeleton graphs serve as the foundation for the GCN-based model, which may investigate as much discriminative information in the temporal and spatial domains as possible.

3.1. Recurrent Neural Networks

RNN is a well-liked neural network design that is often used in natural language processing and voice recognition by taking advantage of temporal correlations between neurons.

Long short-term memory (LSTM) cells, which serve as memory units and aid in capturing long-term dependencies through gradient descent, are often paired with an RNN in HAR.

RNNs for HAR problems have been used in very few publications. The primary issues raised in this research are HAR resource usage and learning pace.

An LSTM-RNN-based HAR framework was developed by Gaur *et al.* (2022). Wearable sensors are included in the framework. Four modules make up the suggested framework: the first deals with data pre-processing; the second discusses the advantages and disadvantages of the RNN model; the third focuses on the LSTM network model; and the fourth is the LSTM and RNN combination module.

An independently recurrent neural network (IndRNN) was introduced by Li *et al.* (2019) to address five RNN issues in HAR. The first addresses the gradient vanishing and exploding problems, even while processing longer sequences with more than 5000 steps. The second enables the construction of deeper networks (more than 20 layers, or even deeper if GPU memory permits). The third allows ReLU to be used for robust training. The fourth enables the interpretation of IndRNN neuron behaviour independently of other influences. The fifth is reduced computational complexity; for lengthy sequences, it is more than ten times faster than cuDNN LSTM.

Study by Gaur and Kumar Dubey (2022); Inoue *et al.* (2018) suggested a model that yields high throughput for HAR after examining a number of model parameters. A binarized BLSTM-RNN model was presented in another paper by Edel and Köppe (2016), in which the inputs, outputs, and weight parameters of all hidden layers are binary values. RNN-based HAR models are primarily concerned with handling resource-constrained settings while maintaining high performance.

Sheng Yu *et al.* (2019) proposed Pseudo Recurrent Residual Neural Networks (P-RRNNs) for video-based human activity recognition. The suggested two-stream P-RRNNs incorporated local spatial and temporal characteristics extracted by a two-stream CNN (GoogLeNet) into global long-term temporal representations. For final categorization, a Softmax layer merged the outputs. After testing on the UCF101 and HMDB51 datasets, the model performed well in human action recognition.

RNN-based methods, especially those using LSTM and BiLSTM units, are well suited for analysing sequential skeletal data because of their capability to model temporal dependencies in variable-length input sequences. They excel in capturing long-range temporal dynamics in action sequences and are computationally efficient, making them ideal for resource-limited platforms. However, they treat skeletal joints as sequential elements, neglecting inter-joint spatial relationships, which can hinder the distinction between actions with similar poses. RNN models also face challenges with scalability, longer training times, and sensitivity to noise and missing frames. While effective in wearable and IoT environments for real-time monitoring, they require additional spatial modelling for tasks requiring detailed pose discrimination.

3.2. Convolutional Neural Networks

This method makes use of CNN's exceptional advantages for object detection in 2D space (image space). In order to learn the spatiotemporal skeleton characteristics, it converts a 3D representation of human skeletons into a 2D array (possibly using spatial relations from the skeleton joints). Figure 7 shows the CNN-based method.

Huynh *et al.* (2019) proposed a CNN-based approach for 3D action recognition that involved extracting pose features from 3D skeleton data and encoding them into images. To enhance diversity, data augmentation techniques were applied to the encoded pose images. Temporal data augmentation was introduced to prevent overfitting and enhance generalisation by generating manifold action images from a single skeleton sequence through the random addition or elimination of some skeleton frames.

Chen *et al.* (2021) suggested a CNN-based method for skeleton-based action recognition using relative features. These features represent human actions and are encoded into colour images. Augmentation techniques such as random scaling, translation, rotation, and noise addition are applied to introduce variations.

Tasnim *et al.* (2020) suggested a DCNN model to train the feature vectors converted from the joint coordinates along the X, Y, and Z axes. With *j* being the joint number, Fi (Xij, Yij, Zij) represents the joints of each i_{th} frame.

For skeletal action recognition, Li and Sun (2021) suggested a CNN fusion approach. Two types of feature vectors, namely three SPI (skeletal posture image) sequences and three STSIs (skeletal trajectory shape images), were used to train the fusion model.

Duan *et al.* (2022) suggested a PoseConv3D model. The spatiotemporal dynamics of skeletal sequences were captured by this model using 3D-CNN. Three-dimensional heatmap volumes serve as the input to the 3D-CNN backbone. The obtained pseudo-heatmaps for limbs and joints are useful inputs for 3D-CNNs.

CNN-based methods for 3D Human Action Recognition (HAR) utilise skeletal joint sequences encoded as 2D pseudo-images, allowing spatial convolutional operations to extract feature patterns. This approach benefits from architectural advancements in image classification and can capture local joint relationships. Data augmentation techniques improve model generalisation, while studies such as PoseConv3D show that 3D convolutional architectures can outperform Graph Convolutional Networks (GCNs) in recognition tasks. However, limitations include potential information loss during encoding, which affects the capture of relational features between joints. Although CNNs scale well with large datasets, high-resolution volumetric representations lead to significant memory requirements. Their applicability is strong when combined with RGB image processing in fields such as surveillance and healthcare, where lightweight architectures can provide a balance between accuracy and inference speed.

3.3. Graph Convolutional Networks

The 3D human skeleton is naturally represented as a graph in GCN-based deep learning, with each joint serving as a vertex and each segment connecting the various body components serving as an edge. The temporal and spatial characteristics of skeletal graph series are often extracted using this method. This method has drawn a great deal of research interest in the last four years due to the benefits of the characteristics that can be derived from the skeleton graph.

Ding *et al.* (2019) proposed a new end-to-end network called AR-GCN (attention-enhanced recurrent graph convolutional network). AR-GCN is an end-to-end network

that can overcome the drawbacks of learning using only key frames and key joints and selectively learn discriminative spatiotemporal information. The benefits of both an RNN and a GCN are combined in AR-GCN. As a result, AR-GCN enhances the GCN's capacity to extract spatial features and model discriminative temporal information.

The Bidirectional Attentive Graph Convolutional Network, or BAGCN, was suggested by Gao *et al.* (2019). Human skeletal sequences are utilised to learn spatiotemporal context using a GCN-based focusing and diffusion process. In order to effectively promote the way messages travel in the graph, the characteristics of BAGCN are constructed around the representation of skeletal data in a single frame by two opposite-direction graphs.

The Sym-GNN (Symbiotic Graph Neural Networks) was suggested by Li *et al.* (2021) for HAR and motion prediction using a 3D human skeleton. Two component networks make up Sym-GNN: a dual bone-based network for learning body-bone-based features and a primary joint-based network for learning body-joint-based features. In essence, each network's backbone is a multi-branch, multi-scale GCN.

A dense connection block was suggested by Wu *et al.* (2019) for ST-GCN in order to increase feature robustness and learn global information. The findings of the suggested approach, which are based on the dense connection block and the spatial residual layer, are superior to those of state-of-the-art techniques that rely on the spatial-temporal GCN.

Sighencea *et al.* (2023) introduced the Spatial-Temporal-GCNN, the first graph-based model for human joint topology. The researchers employed graph convolutional neural networks (GCNNs) to model the spatial characteristics of skeleton data and temporal convolutions to extract motion information from videos, yielding positive results. While GCNNs excel in graph topology modelling, ST-GCNN's manual topology definition hinders the modelling of connections between non-directly connected joints, limiting their expressive ability. In order to optimise GCNN performance, various studies have addressed the following issues.

Huang *et al.* (2023) proposed Channel-Topology Refinement-GCNN to address the limitations of static graphs in representing topological relationships. This transforms static graphs into dynamic ones, allowing more dynamic correlations between non-naturally connected joints and effective channel feature aggregation. Bulugu *et al.* (2024) created Adaptive Structure-GCNN to represent dependencies between non-naturally linked joints, replacing fixed adjacency matrices with learnable ones for better spatial topological information.

Tasweer Ahmad *et al.* (2021) offered the first thorough analysis of graph convolutional network (GCN) techniques for human action identification based on skeletons. Because the skeleton of the human body can be represented as a graph, GCNs have become effective tools for modelling non-Euclidean data and have greatly improved the effectiveness of action detection. Along with analysing benchmark datasets and compiling pertinent resources and open-source programs, the study offered a taxonomy of GCN approaches. Along with outlining future research goals and highlighting present obstacles, it gave academics a systematic understanding of GCN-based methods for human action identification.

Graph Convolutional Networks (GCNs) are the standard for skeleton-based action recognition, effectively modelling human body structure by treating joints as nodes and bones as edges. They adeptly encode spatial and temporal dependencies and can learn dynamic topologies, capturing long-range interactions between joints. Attention mechanisms enhance feature extraction, achieving over 92% accuracy on the NTU RGB+D 120 benchmark. However, GCNs face limitations such as reliance on static topologies, quadratic computational complexity with node counts, and poor performance in uncontrolled environments or outdoor settings. GCNs are best suited for controlled applications such as clinical monitoring, while their scalability and adaptability to real-time systems pose significant challenges.

3.4. Hybrid-DNN

Deep learning networks are used in hybrid-DNN techniques to train recognition models and extract features. The Attention Enhanced Graph Convolutional LSTM Network, or AGC-LSTM, was suggested by (Si *et al.*, 2019). Combining discriminative characteristics in temporal dynamics and spatial configuration, as well as investigating the co-occurrence connection between temporal and spatial domains, are all possible with AGC-LSTM. Additionally, the AGC-LSTM layer greatly lowers computational costs by learning high-level semantic representations.

In order to create the end-to-end trainable framework, a Bayesian neural network (BNN) model is once again merged with LSTM and graph convolution (Zhao *et al.*, 2019). The temporal dependence of posture changes over time is captured by LSTM, while the spatial dependency across various body joints is captured by graph convolution.

The end-to-end Semantics-Guided Neural Network (SGN) is suggested by Zhang *et al.* (2020) by combining CNN and GCN models. It is composed of two modules: one at the joint level and one at the frame level. SGN combines a joint's location and velocity data to learn the joint's dynamics representation. SGN uses two CNN layers and three GCN layers, respectively, to represent the interdependence of frames and joints in the joint-level module.

Using information from the human skeleton, Trivedi *et al.* (2022) suggested PSUMNet (Part Stream Unified Modality Network) for HAR. In contrast to traditional modality-wise streaming methodologies, it presents the combined modality part-based streaming strategy. Compared to state-of-the-art techniques, PSUMNet performs well across skeletal action recognition datasets while reducing the number of parameters by around 100–400%.

Hyperformer is a hybrid model that was constructed by Zhou *et al.* (2022). This model used a graph distance embedding technique to provide bone connections to the Transformer. Hyperformer keeps the skeletal structure throughout training, in contrast to GCN, which only utilises it for initialisation. Additionally, Hyperformer incorporates inherent higher-order interactions into the model by implementing a self-attention (SA) mechanism on hypergraphs called Hypergraph Self-Attention (HyperSA).

Bavil *et al.* (2023) proposed the Action Capsule Network (CapsNet) for skeleton-based action recognition. ResTCN (Residual Temporal Convolution Neural Network) and CapsNet are used to hierarchically encode the temporal information linked to each joint in order to dynamically concentrate on a subset of crucial joints. CapsNet learns to dynamically pay attention to critical joint characteristics and every movement of the human skeleton.

Nguyen *et al.* (2023) reviewed deep learning methods for human activity recognition (HAR) utilising 3D human skeletal data. It examined RNN, CNN, GCN, and Hybrid-DNN models, datasets, metrics, and outcomes from 2019 to March 2023. The KLHA3D-102 and KLYOGA3D datasets were compared to show the performance of CNN-, GCN-, and Hybrid-DNN models. The review revealed existing successes and issues, helping to choose solutions and create realistic HAR applications.

Nadeem *et al.* (2024) proposed Hybrid-Graformer, a deep learning model for skeleton-based multi-person, multi-view human action recognition that combines a partitioning Transformer with a Graph Convolutional Network (GCN) backbone. Whereas the Transformer combines short- and long-term temporal dependencies using an effective partitioning technique, the attention-based GCN captures intrinsic skeletal topology and discriminative properties. To increase robustness, a new cosine-based noise augmentation technique for joints across time was also presented. On a number of benchmark datasets, the experimental findings demonstrated that Hybrid-Graformer attained accuracy on par with state-of-the-art techniques.

Hybrid deep neural network architectures, combining RNNs, CNNs, GCNs, and Transformers, achieve state-of-the-art results in 3D HAR benchmarks by optimising spatial and temporal feature extraction. Advantages include reduced computational costs and improved semantic representation through models such as AGC-LSTM, which merges GCN and LSTM techniques, and SGN, which adds high-level semantic guidance. However, the complexity and increased parameter count of hybrid models present limitations, such as longer training times and challenges in cross-dataset generalisation. Scalability is also a concern, as these models are computationally intensive, although PSUMNet achieves parameter reductions while retaining accuracy. For real-world scenarios, hybrid architectures excel in applications requiring high accuracy but may necessitate model compression techniques for efficient deployment.

3.5. Transformer-Based Methods

The vision community closely monitors Transformers (Vaswani, 2017) and extends them to vision tasks such as image classification (Dosovitskiy *et al.*, 2020; Ren *et al.*, 2023; Touvron *et al.*, 2021), object detection (Carion *et al.*, 2020; Zhu *et al.*, 2020), segmentation (Ye *et al.*, 2019), image restoration (Ma *et al.*, 2024; Li *et al.*, 2023), and point cloud registration (Huang *et al.*, 2023; Mei *et al.*, 2023; Wang *et al.*, 2023). Transformers (Vaswani, 2017) have proven their overwhelming power on a wide range of language tasks (e.g., text classification, machine translation, or question answering) (Khan *et al.*, 2022). Point-centric research has undergone a significant change with the advent of Transformer algorithms. By demonstrating encouraging improvements in accuracy and computational efficiency, these Transformer-based techniques are progressively undermining the supremacy of GCN approaches. Based on our study, we are certain that Transformer-based methods have a great deal of promise and will soon become the standard method.

The core module in the Transformer, multi-head self-attention (MSA) (Dosovitskiy *et al.*, 2020), aggregates sequential tokens with normalized attention as:

$$z_j = \Sigma_i \ \text{softmax}\left(\frac{QK}{\sqrt{d}}\right)_i V_{i,j}$$

where Q, K, and V are query, key, and value matrices, respectively. d is the dimension of the query and key, and (z_j) is the jth output token. This step usually represents the computation and update of the contextual relationships of the overall 3D skeleton features. Building upon the MSA mechanism of the Transformer for solving the 3D-SAR problem, many Transformer architecture-based solutions have been proposed.

Cho *et al.* (2020) introduced the Self-Attention Network (SAN) model, which fully employs the self-attention mechanism to represent spatiotemporal correlations. Shi *et al.* (2020) introduced DSTA-Net, a decoupled spatiotemporal attention network

with decoupled position encoding and global spatial regularization. DSTA-Net divides skeletal data into four streams: spatiotemporal, spatial, slow-temporal, and fast-temporal. Each stream focuses on a certain component of the activity. A new Spatial-Temporal Transformer network (ST-TR) proposed by Plizzari *et al.* (2021) uses spatial and temporal self-attention modules to capture correlations and dynamic relationships between nodes across frames. To effectively manage action sequences of various durations, Ibh *et al.* (2023) suggested TemPose, which removes padded temporal and interaction tokens from the attention map. TemPose predicts the action class by combining the player and ball positions at the same time.

The Transformer-based method addresses the limitation of focusing on local information and excels at capturing dependencies over extended sequences. The Transformer architecture excels in capturing temporal correlations in skeleton-based human behaviour detection tasks.

The rich high-dimensional semantic information in skeleton data is difficult to capture and encode, limiting its spatial connection modelling (Xiang *et al.*, 2023; Zhu *et al.*, 2023). Several hybrid designs have arisen by combining the Transformer with GCNs or CNNs. These models aim to capitalize on the capabilities of each underlying architecture. Hybrid models combine the Transformer's capabilities with those of RNNs, CNNs, or GCNs to provide a more robust framework for various applications (Yuan *et al.*, 2022; Zhang *et al.*, 2022).

4. Benchmark Datasets and Evaluation Metrics

The performance of deep learning models for HAR based on 3D human skeletal data must typically be assessed using a few benchmark datasets. A few significant databases that include 3D human skeleton data are listed below.

The UTKinect-Action3D dataset (Xia *et al.*, 2012) consists of three different kinds of data: 3D human skeletons, depth images, and colour images. Ten human action types—walking, sitting, standing, picking up, carrying, throwing, pushing, pulling, waving, and clapping—are recorded using a single MS Kinect with the Kinect for Windows SDK Beta Version. Every frame's skeleton data contains 20 joints, each of which includes coordinates (x, y, z), for a total of 199 motion sequences.

The SBU-Kinect dataset (Yun *et al.*, 2012) is obtained from the Microsoft Kinect sensor. The dataset comprises 282 video sequences categorised into eight action classes: "approaching," "departing," "pushing," "kicking," "punching," "exchanging objects," "hugging," and "hand shaking." Each skeleton frame annotates 15 joints per individual using OpenNI with NITE middleware from PrimeSense, and each frame contains two individuals.

The Florence 3D Actions dataset (Seidenari *et al.*, 2013) was recorded with MS Kinect cameras at the University of Florence in 2012. This dataset comprises nine activity classes: wave, drink from a bottle, answer the phone, applaud, clench, sit down, stand up, read, watch, and bow. The activities were executed two to three times by 10 participants, yielding 215 action sequences. Each skeletal frame has 15 body joints, each defined by x, y, and z coordinates and recorded using MS Kinect.

The J-HMDB dataset (Jhuang *et al.*, 2013) is a subset of HMDB (Kuehne *et al.*, 2011) with 21 action categories: brush hair, catch, clap, climb stairs, golf, jump, kickball, pick, pour, pull-up, push, run, shoot ball, shoot a bow, fire a rifle, sit, stand, swing baseball, toss, walk, and wave. The skeletal frame has 15 joints, including 13 joints (left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, and right ankle) and two landmarks (face and abdomen). J-HMDB comprises 928 samples and employs three train/test splits in a 7:3 ratio (70% for training and 30% for testing).

The Northwestern UCLA Multiview Action 3D (N-UCLA) dataset (Wang *et al.*, 2014) was recorded with the MS Kinect Version 1 sensor from multiple perspectives. The training data are obtained from View 1 and View 2, whereas the testing data are sourced from View 3. This dataset has ten activity categories: one-handed pickup, two-handed pickup, garbage disposal, ambulation, sitting, standing, donning, doffing, tossing, and carrying. Each action is executed by 10 subjects.

The SYSU 3D Human-Object Interaction dataset (Hu *et al.*, 2015) comprises 480 skeletal sequences containing 12 action categories executed by 40 distinct individuals. The human skeleton comprises 20 joints. In every action, each subject may interact with one of six objects: a phone, chair, bag, wallet, mop, or besom. The authors used two data-splitting techniques for training and assessment. The first procedure allocates half of the samples for training and the remaining half for testing. The second procedure uses half of the individuals for training and the remaining half for testing.

The NTU RGB+D dataset (Shahroudy *et al.*, 2016) was acquired using three MS Kinect V2 sensors. The 3D skeletal data comprise the three-dimensional coordinates (x, y, z) of 25 principal body joints in each frame. It comprises 56,880 skeletal videos across 60 activity categories. This dataset is divided into two categories: cross-subject and cross-view. The cross-subject dataset comprises 40,320 videos from 20 participants for training, with the remainder allocated for testing. The cross-view setting comprises 37,920 videos recorded from Cameras 2 and 3 for training, whereas those from Camera 1 are used for testing.

The Kinetics-Skeleton dataset (Kay *et al.*, 2017) derives its name from the DeepMind Kinetics human movement video dataset. It contains 400 human activity categories derived from around 300 videos. Each video lasts approximately 10 seconds, and the 3D human skeleton is annotated using the OpenPose toolkit (Cao *et al.*, 2017). Every human skeleton comprises 18 joints. The training and test sets of this dataset include 240,436 and 19,794 samples, respectively.

The NTU RGB+D 120 dataset is an extension of the NTU RGB+D dataset. Most descriptors for this data*et al*ign with those of the NTU RGB+D dataset (Shahroudy *et al.*, 2016), differing primarily in that it comprises 120 action classes and has an increased sample size of 114,480. This dataset is divided into two components: an auxiliary set and a one-shot evaluation set. The auxiliary set comprises 100 classes, and all examples from these classes are available for training. The evaluation set comprises 20 unique classes, with one sample selected from each class as an example, while the remaining samples from these classes are used to assess recognition performance (Kumar *et al.*, 2021). Two evaluation methods for the NTU RGB+D 120 dataset are configured identically to those in the NTU RGB+D dataset, with Cross-Subject retaining the same designation in both datasets, while Cross-View in the NTU RGB+D dataset is renamed Cross-Setting.

The KLHA3D-102 dataset (Kishore *et al.*, 2018) is gathered and aggregated from the MOCAP system's eight cameras. As shown in Figure 13, the 3D human skeleton in each frame contains 39 joints. With 102 classes and five subjects for each action class, KLHA3D-102 contains 510 video sequences (299,468 frames).

The KLYOGA3D dataset has a joint structure similar to that of the KLHA3D-102 dataset (Kishore *et al.*, 2018). The only difference is that it contains 39 yoga action classes; therefore, the total number of video sequences is 39, with 173,060 frames.

The KTH, Weizmann, and IXMAS datasets were used to test the Spiking Separated Spatial and Temporal Convolutions (S3TCs) network (El-Assal *et al.*, 2024). Ten basic action classes performed by nine people make up the Weizmann dataset, whereas 25 subjects perform six types of human behaviours numerous times in four settings in the KTH dataset. To enable multi-view action identification, the IXMAS dataset comprises more everyday activities from several camera viewpoints. By factorizing 3D spiking convolutions into spatial and temporal convolutions, S3TCs identified spatiotemporal information from videos in all three datasets. This reduced the number of parameters, increased spiking activity, and outperformed conventional 3D convolutional techniques while preserving the efficiency of neuromorphic hardware implementation.

Benchmark video and robotic-assisted surgery datasets were analysed by Grid Keypoint Learning (Gao *et al.*, 2021). Condensation loss and a 2D binary map identify keypoints in discrete grid space for reliable and meaningful feature extraction. The model retains spatial structure for long-term video prediction by stochastically propagating grid keypoints. Experiments revealed that it surpassed state-of-the-art approaches and reduced computation by 98%, offering potential surgical applications.

4.1. Evaluation Metrics

The metrics are crucial and need to be consistent in order to assess and compare the performance of HAR models based on 3D human skeletons. In machine learning, the accuracy metric is often used to evaluate models as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

where the number of predictions that are true when the label is positive is called TP (True Positive), and the number of predictions that are true when the label is negative is called TN (True Negative). The number of predictions with a positive label but a false prediction is known as FP (False Positive), while the number of predictions with a negative label but a false prediction is known as FN (False Negative).

A thorough overview of current deep learning-based techniques for 3D HAR using skeletal data is given in Table 3. The table lists some cutting-edge methods, emphasizing their primary objectives, the datasets used, algorithmic frameworks, the accuracy attained, and their related limitations.

Table 3: Survey of 3D Human Action Recognition Based on Deep Learning Methods

Ref	Main Objective	Dataset	Algorithms	Result (Accuracy)	Limitation
Li *et al.* (2019)	To improve human action recognition using IndRNN for better long-term dependency learning.	NTU RGB+D 120 dataset	DenseIndRNN	93.7	Negative recurrent weights may cause oscillations and remain difficult to interpret.

Ding *et al.* (2020)	To enhance skeleton-based action recognition.	NTU RGB+D	Semantics-Guided GCNs	86.2	Limited receptive fields and uniform data usage hinder full global context understanding.
Yang *et al.* (2020)	To improve human action recognition by integrating multiple graph embedding features for richer skeleton representation.	NTU RGB+D, Kinetics, and SYSU-3D	Graph Convolutional Network with unified graph embeddings	90.3	Limited temporal correlation modelling and higher computational complexity.
Zhang *et al.* (2019)	To enable fine-grained input modulation and improve performance on sequential tasks.	N-UCLA dataset	EleAtt-RNN (RNN, LSTM, GRU with EleAttG)	80.7	Increased model complexity and possible computational overhead due to the attention mechanism.
Li *et al.* (2018)	To develop a robust method to analyse and classify complex human motion.	MSR Action3D, UTD MHAD, SYSU, and NTU RGB+D datasets	Bayesian GC-LSTM	81.8	High computational complexity limits real-time use.
Chen *et al.* (2021)	To improve skeleton-based action recognition by learning and adapting feature aggregation graph topologies.	NTU RGB+D, NTU RGB+D 120, and NW-UCLA datasets	Channel-wise Topology Refinement Graph Convolution (CTR-GC)	92.4	Limited ability to capture long-range temporal dependencies beyond graph topology refinement.
Trivedi and Sarvadevabhatla (2022)	To develop a scalable and efficient approach for pose-based action recognition.	NTU RGB+D 60/120, NTU 60-X/120-X, and SHREC	PSUMNet (Part Stream Unified Modality Network)	89.4	High parameter usage in prior multi-stream models.
Zhang and Li (2024)	To enhance the recognition of fine-grained skeletal motions.	NTU RGB+D and NTU RGB+D 120	Skeleton-based Linkage Graph Convolutional Neural Network	92.49	Contrastive learning overhead may hinder real-time deployment.

5. Optimization in Deep Learning-Based HAR

Several researchers have suggested optimization-enhanced frameworks to address the drawbacks of traditional deep learning-based Human Action Recognition (HAR) techniques, including their high computational cost, poor generalisation across datasets, and inefficiencies in hyperparameter tuning. These optimised methods combine deep learning architectures with metaheuristic algorithms to increase real-time applicability, decrease overfitting, and improve accuracy. The following section examines recent optimisation methods used with DL-based HAR systems, highlighting how they have helped enhance performance and overcome earlier drawbacks.

The optimal DL-based HAR approach described by Al-Wesabi *et al.* (2021) used MobileNet-v2 as a feature extractor and BiLSTM as a classifier. The Chaos Game Optimisation (CGO) technique is used to optimise BiLSTM model hyperparameters to enhance recognition performance. The novel aspect of this study is the use of the CGO algorithm to optimise HAR hyperparameters. To evaluate the effectiveness of the proposed ODL-HAR approach, several simulations were performed using two benchmark datasets. Experimental findings show that the ODL-HAR method outperforms other HAR techniques across many assessment criteria.

The study by Basak *et al.* (2022) proposed DSwarm-Net, a deep learning framework for HAR that uses 3D skeleton data for action classification. Four features are extracted from the data: Distance, Distance Velocity, Angle, and Angle Velocity. These features are encoded into images, making the classification task less complex. The model was trained on three publicly available HAR datasets, achieving competitive results and demonstrating superiority compared to state-of-the-art models.

Challa *et al.* (2023) aimed to develop an optimized Deep Learning (DL) model for recognizing human activities captured through IMU sensors. The model combines convolutional neural network (CNN) layers and bidirectional long short-term memory (Bi-LSTM) units to extract spatial and temporal sequence features from raw sensor data. The model uses the Rao-3 metaheuristic optimization algorithm to identify ideal hyperparameter values. Validated on the PAMAP2, UCI-HAR, and MHEALTH datasets, the proposed DL model achieved accuracies of 94.91%, 97.16%, and 99.25%, respectively, outperforming existing state-of-the-art models.

To train the model on sequential and spatial patterns, Thakur *et al.* (2025) used Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) techniques. The use of optimisation methods has improved the attribute selection process. This research introduces a hybrid optimisation technique that combines the Whale Optimisation Algorithm with the Grey Wolf Optimiser. The hybrid strategy improves the overall optimisation process by integrating and utilising the advantages of both techniques. Additionally, this study could help people who have had strokes or amputations choose a suitable model and create a customised gait reference trajectory.

Modified Wild Horse Optimisation with DL-Aided Symmetric Human Activity Recognition (MWHODL-SHAR) was developed by Tahir *et al.* (2023). The main goal of the MWHODL-SHAR model is to recognise symmetric activities such as running, walking, standing, and sitting. With the MWHODL-SHAR approach, human activity data are pre-processed in phases to facilitate subsequent processing. To recognise activities, a CNN-ALSTM model based on a convolutional neural network is used. The MWHO method is employed to optimise CNN-ALSTM detection rates through hyperparameter adjustment. A benchmark dataset is used to perform the experimental validation of the MWHODL-SHAR approach. A comprehensive comparative analysis showed that the MWHODL-SHAR strategy outperformed other current techniques.

Ahmad *et al.* (2024) used a "1D Convolutional Neural Network (1DCNN)" and a "Gated Recurrent Unit (GRU)" to create an Adaptive Hybrid Deep Attentive Network (AHDAN) for Human Activity Recognition. Enhanced Archerfish Hunting Optimiser (EAHO) is recommended to fine-tune network settings for better recognition. To validate the HAR model, deep learning networks and heuristic methods are tested. EAHO-based HAR models surpass deep learning networks in recall, specificity, and precision, with accuracy values of 95.36%, 95.25%, 95.48%, and 95.47%, respectively. The findings showed that the model can recognise human motion quickly. Optimisation decreases computational complexity and overfitting.

Amirthalingam *et al.* (2024) created a sensor-based hand gesture recognition system to predict medicine use. This study develops a smart sensor device-based hand gesture prediction algorithm for medicine ingestion. The device measures hand motions using a tri-axial gyroscope, geometry, and accelerometer. Smartphone software collects hand gesture data from the sensor device and stores them in a .csv cloud database. The novel Sea Horse Optimization–Deep Neural Network model collects, processes, and classifies these data to detect medicine consumption activities. The DNN model's biases, weights, and hidden layers are updated using SHO. The updated parameters help the DNN model categorise hand motion data to detect pharmaceutical activities.

Gupta *et al.* (2021) introduced CNN-GRU, a hybrid deep neural network model for Human Activity Recognition (HAR), utilizing accelerometer and gyroscope inputs from smartphones and smartwatches. Deep learning automatically extracts optimal features and finds hidden patterns in raw sensor data, unlike laborious feature engineering. The CNN-GRU model uses convolutional layers for spatial feature learning and gated recurrent units for temporal dependencies. The model outperformed InceptionTime and DeepConvLSTM created by AutoML on the WISDM dataset, making it suitable for biometrics and remote health monitoring.

Raziani *et al.* (2021) evaluated wearable and smartphone sensor data for Human Activity Recognition (HAR) categorization in healthcare, senior care, rehabilitation, and smart surveillance. Online HAR data categorization used a one-dimensional CNN and seven metaheuristic algorithms to automatically adjust its hyperparameters. The technique was tested on the UCI HAR dataset against cutting-edge evolutionary algorithms and deep learning models. Experiments showed that metaheuristic-based optimization made the CNN model for HAR more robust and efficient.

5.1. Comparative Discussion of Optimization Algorithms

Metaheuristic optimization algorithms have emerged as essential tools for automating hyperparameter tuning in deep learning-based HAR systems, addressing the prohibitive cost of exhaustive grid search across high-dimensional parameter spaces. This section provides a critical comparative analysis of the key optimization algorithms reviewed in this survey, highlighting their mechanisms, relative advantages, limitations, and the HAR scenarios for which each is most appropriate. Table 4 below summarizes the comparison across seven algorithms reviewed in this survey.

Table 4: Comparative Analysis of Metaheuristic Optimization Algorithms Applied in HAR Deep Learning

Algorithm	**Mechanism**	**Advantages**	**Limitations**	**Best Use in HAR**
Grey Wolf Optimizer (GWO)	Mimics grey wolf hunting hierarchy (alpha, beta, delta, omega)	Strong exploration–exploitation balance; fast convergence; fewer parameters	May converge prematurely in complex high-dimensional spaces	Tuning LSTM/GCN layer counts and learning rates where convergence speed matters
Whale Optimization Algorithm (WOA)	Simulates bubble-net hunting of humpback whales	Effective global search; robust against local optima	Computationally expensive; sensitive to initial population	Hyperparameter search in CNN-based HAR with large feature spaces
Chaos Game Optimization (CGO)	Uses fractal-based chaotic maps for population initialization	Avoids premature convergence; diverse initial solutions	Complex implementation; limited theoretical analysis	Hybrid DNN architectures with multimodal HAR data
Particle Swarm Optimization (PSO)	Simulates social behavior of bird flocking	Simple, fast, widely validated in deep learning tasks	Prone to local optima in high-dimensional search spaces	Initial hyperparameter search for RNN and LSTM-based HAR models
Sea Horse Optimization (SHO)	Mimics sea horse foraging and mating strategies	Promising results on sensor-based HAR; good diversity	Relatively new; limited domain-specific validation for 3D HAR	Wearable sensor-based gesture/activity recognition tasks
Modified Wild Horse Optimization (MWHO)	Inspired by wild horse herd movement and leadership	Good balance for symmetric activity classification	Needs domain tuning; limited benchmark comparisons	Symmetric activity recognition with CNN–ALSTM architectures
Rao-3 Algorithm	Utilizes best–worst solution interaction without algorithm-specific parameters	Parameter-free; low computational overhead	May require more iterations for high-dimensional problems	IMU sensor-based HAR where computational budget is limited

Grey Wolf Optimizer (GWO) is the most suitable choice when rapid convergence is required, and the hyperparameter space is moderately complex. Its hierarchical hunting mechanism provides a reliable exploration–exploitation balance, making it effective for tuning LSTM layer configurations and GCN adjacency parameters. Studies reviewed in this survey confirm that GWO–WOA hybrid approaches outperform either algorithm alone on wearable sensor-based HAR datasets.

The Whale Optimization Algorithm (WOA) is preferred for problems involving large feature search spaces, such as multi-stream GCN architectures, where the interaction between parameters is non-linear and complex. Its bubble-net search strategy performs effective global exploration in the early optimization phase, though its computational cost increases substantially with population size and dimensionality (Lale *et al.*, 2026).

Chaos Game Optimization (CGO) is particularly effective when the hyperparameter landscape is multi-modal with many local optima. Its fractal-based initialization produces diverse starting solutions that reduce the risk of premature convergence. CGO-optimized BiLSTM models outperform manually tuned baselines on multiple HAR benchmark datasets, suggesting strong applicability for complex hybrid architectures.

Particle Swarm Optimization (PSO) remains the most computationally efficient choice for initial hyperparameter exploration, particularly for RNN-based models with fewer tunable parameters. Its simplicity and low overhead make it suitable for scenarios with limited computational budgets, though it is prone to premature convergence in high-dimensional spaces.

Sea Horse Optimization (SHO) shows promise for wearable sensor-based HAR applications, as demonstrated by (Amirthalingam *et al.*, 2024) in hand gesture recognition. However, it remains relatively untested on 3D skeletal HAR benchmarks, and further validation on large-scale datasets is needed before broader adoption.

A key observation from this review is that no single optimization algorithm consistently outperforms all others across all HAR tasks. Hybrid optimization strategies, such as GWO–WOA combinations, provide a practical solution by leveraging the complementary strengths of individual algorithms. Future research should focus on developing HAR-specific adaptive optimization frameworks that dynamically select and combine algorithms based on dataset characteristics and architecture complexity.

5.2. Limitations and Scalability Concerns in Optimization-Enhanced HAR

- **Computational Overhead:** Metaheuristic optimization requires multiple full model training runs to evaluate candidate hyperparameter configurations, multiplying the total training cost by the population size and number of iterations. For large-scale datasets such as NTU RGB+D 120, this can be prohibitive without distributed computing infrastructure.
- **Cross-Dataset Generalization:** Optimal hyperparameters identified on one dataset frequently do not transfer to others. This is a critical limitation for real-world HAR deployment, where diverse data distributions must be handled. Transfer learning combined with lightweight fine-tuning via optimization is a promising direction to address this gap.
- **Reproducibility:** Most reviewed optimization studies report results on single train-test splits without statistical significance testing, making it difficult to assess the true contribution of optimization versus dataset-specific effects. Future studies should report results over multiple runs with variance analysis to ensure reproducibility.

6. Applications of HAR

The many application areas where HAR has been successfully implemented are shown in Table 5. It lists the precise tasks that HAR systems carry out in each area. It also summarizes the present performance levels achieved, with accuracy rates, especially in security and healthcare applications, often above 90%. In a variety of sectors, this illustrates how important HAR is for enhancing safety, efficiency, and personalized experiences.

Table 5: Applications of Human Activity Recognition

Application Domain	Tasks	Current Performance
Healthcare (Wan *et al.*, 2020; Chen *et al.*, 2019)	Fall detection, monitoring patient movement, physical therapy assistance, chronic disease management, sleep monitoring, elderly care, rehabilitation support	Highly accurate fall detection systems with 95%+ accuracy; real-time movement monitoring in hospitals and elderly care; effective patient monitoring and rehabilitation progress tracking; personalized treatment adjustments for chronic disease management
Fitness and Sports (Ramasamy Ramamurthy & Roy, 2018; Jegham *et al.*, 2020; Nadeem *et al.*, 2021; Khan *et al.*, 2024)	Activity tracking and evaluation, technique improvement, injury prevention, sports performance analysis, virtual coaching	Accurate activity tracking with 90%+ accuracy; enhanced feedback for sports technique improvement and injury prevention; real-time performance analysis for professional athletes
Security (Dhiman & Vishwakarma, 2019; Beddiar *et al.*, 2020; Shreyas *et al.*, 2020; Daga & Meena, 2022)	Surveillance and suspicious activity detection, access control, intrusion detection, biometric authentication, public safety (traffic monitoring, crowd management)	Surveillance systems achieving over 90% accuracy in detecting abnormal behavior; biometric authentication with accuracy rates surpassing traditional methods; crowd management solutions enhancing public safety in high-traffic areas
Smart Homes (Dhiman & Vishwakarma, 2019; Beddiar *et al.*, 2020; Shreyas *et al.*, 2020; Daga & Meena, 2022; Anbazhagan *et al.*, 2024)	Energy optimization, automation (light control and curtain operation), personalized environment settings, security (identifying individuals, detecting unusual behaviors)	Automation systems achieving 80%–90% energy efficiency gains; improved security and customization through activity-based controls; effective behavior detection contributing to personalized user experiences
Virtual and Augmented Reality (D'Sa & Prasad, 2019; Coronado *et al.*, 2023)	VR/AR experiences enhanced by movement feedback, virtual training and education, personalized virtual interior design, marketing applications (tailored content based on activity)	Highly responsive movement-based feedback systems; immersive virtual learning environments with over 85% accuracy in interaction detection; increased user engagement in marketing with personalized content

7. Open Challenges and Future Research Directions

Accurately identifying human behaviors requires addressing a number of challenges associated with the activity recognition process. The execution of multiple activities simultaneously is one of the most significant obstacles, which can make recognition difficult (D'Sa & Prasad, 2019). Furthermore, weather conditions and lighting (Beddiar *et al.*, 2020), as well as video contrast, may affect how effectively activities are recognized in different environments. During processing, self-occlusion, object occlusion, and shadows may result in imprecise tracking and detection (Gilroy *et al.*, 2021). Another difficulty is scalability (Hossain *et al.*, 2018), as recognition accuracy may be affected by the distance between the camera and the activity.

Although the camera itself is essential for activity detection (Paul *et al.*, 2013), its specialized features and resolution may make it challenging to accurately identify activities. Additionally, moving background objects or people may cause issues, as they can be confused with actual objects or people, leading to inaccurate activity identification (Poppe, 2010). Another issue is the scarcity of sufficient data (Chaquet *et al.*, 2013), as obtaining labeled data for HAR can be costly and time-consuming.

The accuracy of recognition may also be affected by intra-class variability (Cherla *et al.*, 2008) and inter-class similarity (Subetha & Chitrakala, 2016), as different individuals may perform the same task in different ways. Therefore, certain human behaviors may appear similar to one another. Model complexity is another issue, as developing DL models for HAR can be computationally expensive and resource-intensive, and their complexity may reduce practicality in real-world scenarios.

Finally, transferability is a major concern (Ghadiyaram *et al.*, 2019), as DL algorithms trained on a particular dataset may not generalize well to other settings or populations, requiring careful consideration of the characteristics of both training and testing datasets.

8. Outcome of the Survey

The results of this extensive survey show several significant trends and unresolved issues in optimization-based deep learning for 3D Human Action Recognition. Critical review of outcomes in the studies examined suggests that although benchmark accuracies have increased considerably, with both leading GCN and hybrid models achieving more than 92% on NTU RGB+D 120, this progress does not necessarily translate to real-world deployment performance.

Recent literature highlights an increasing recognition that benchmark saturation on NTU datasets does not necessarily translate to performance in more challenging real-world conditions. Architectures tested on the ANUBIS benchmark, a more recently introduced dataset featuring rear-view occlusions and fine-grained social interactions, perform 30–50 times worse than NTU scores, even with state-of-the-art GCN approaches (Liu *et al.*, 2022). This indicates the strong need for evaluation protocols that stress-test models under realistic conditions such as viewpoint variation, partial occlusion, and cluttered backgrounds.

Metaheuristic optimization has also been shown to enhance the performance of HAR models, particularly for hyperparameter-sensitive architectures such as BiLSTM and CNN-ALSTM. However, not all studies provide actionable conclusions due to the absence of a unified experimental framework for comparing different optimizers under identical conditions. Recent research using multiple metaheuristic algorithms to optimize CNN and Transformer models for sequence forecasting reported that GWO-optimized models achieved the lowest error rates, whereas WOA was computationally more expensive with only marginal accuracy gains (Lale *et al.*, 2026). Such comparative evidence would be valuable for the HAR domain.

From an application perspective, the reviewed literature confirms that healthcare and rehabilitation remain the most advanced domains for HAR deployment, with fall detection systems achieving above 95% accuracy in controlled environments (Chen *et al.*, 2019; Wan *et al.*, 2020). Performance in constrained settings is comparable in security and surveillance applications; however, robustness under crowded scenes, lighting variations, and viewpoint ambiguity remains a key research challenge (Dhiman & Vishwakarma, 2019; Beddiar *et al.*, 2020). Human–computer interaction and smart home applications introduce additional complexity due to multi-person interactions and fine-grained activity variations, where hybrid and transformer-based models show strong potential.

One of the most promising directions identified in this survey is two-stream spatiotemporal GCN–Transformer architectures. Chen *et al.* (2025) introduced a parallel GCN–Transformer (SA-TDGFormer) that jointly learns local structural features using GCNs and global temporal dependencies using self-attention. This parallel integration approach avoids the information bottleneck commonly observed in sequential hybrid designs and warrants further investigation for real-time HAR applications.

The importance of dataset diversity in advancing HAR research cannot be overstated. Although the NTU RGB+D family remains the most widely used benchmark, its indoor and laboratory-controlled nature limits the ecological validity of reported results. Systematic evaluation across diverse datasets, including indoor and outdoor settings, varying camera viewpoints, different skeletal capture conditions,

and multi-person scenarios, should be prioritized in future work. The ANUBIS benchmark and KLHA3D-102 dataset are important contributions toward this objective.

Finally, a key gap is the limited interpretability of deep learning models in HAR. Since HAR systems are increasingly used in safety-critical applications such as autonomous driving and healthcare monitoring, the lack of explainability limits regulatory acceptance and user trust. Integrating post-hoc explainability methods, including gradient-based attention visualization and SHAP-based analysis, with GCN and hybrid architectures represents an important future research direction (Liu *et al.*, 2022). As shown in Table 6, spatiotemporal feature refinement, skeleton-based semantic modeling, and multi-view learning achieve state-of-the-art or near state-of-the-art performance in terms of accuracy, generalization, and efficiency, indicating strong potential for real-world deployment in complex HAR scenarios.

Table 4: Studies on Optimization-Driven Deep Learning Models For 3D Human Action Recognition (Single-Person and Group Activities)

Dataset	**Model / Method**	**Activity Type**	**Novel Aspects**	**Results / Accuracy**
NTU RGB+D	Pose estimation maps (Liu & Yuan, 2018)	Single-person (3D)	Uses pose estimation maps as richer cues for body part movements, robust to noisy 2D poses.	91.71%, 95.26%
NTU RGB+D	RNN (mfRNN) (Wang *et al.*, 2020)	Single-person (3D)	Novel recurrent neural network to handle unobserved video frames.	75.40%, 71.04%, 66.02%
IXMAS / nvGesture	RNN (mfRNN) (Wang *et al.*, 2020)	Single-person (3D)	Optimized recurrent learning across missing or noisy frames.	~97% accuracy
NTU-60 / NTU-120 / SYSU	SGN (Zhang *et al.*, 2020)	Skeleton-based (3D)	Semantics-guided network optimizing skeleton-based 3D recognition.	2.1%–1.7% improvement (CS & CV)
MHAD / UWA / DHA	GMVAR (Wang *et al.*, 2019)	Multi-view 3D	Combines adversarial training with View Correlation Discovery Network (VCDN) to optimize multi-view recognition.	98.94%, 76.28%, 88.72%
Charades / MultiTHUMOS	Coarse-Fine Networks (Kahatapitiya & Ryoo, 2021)	Single-person (3D video)	Dynamically selects temporal frames for multi-scale representation.	25.10% mAP
Volleyball / Collective Activity	ARG (Actor Relation Graph) (Wu *et al.*, 2019)	Group	Graph-based optimization of player relationships in multi-person scenes.	Volleyball: 92.6%, Collective Activity: 91.0%
Volleyball / Collective Activity	GroupFormer (Li *et al.*, 2021)	Group	Spatial-temporal contextual optimization with Transformer-based refinement.	Volleyball: 95.7%, Collective Activity: 96.3%
Volleyball	Actor-Transformer (Gavrilyuk *et al.*, 2020)	Group	Transformer-based network for actor-level refinement in group activity.	94.40%
Volleyball	Skeleton-based method (Zappardino *et al.*, 2021)	Group (skeleton)	State-of-the-art group activity recognition using only skeleton data.	90%
CAD / VD	GLIL (Graph LSTM-in-LSTM) (Shu *et al.*, 2021)	Group / Sports	Hierarchical modeling of person-level and group-level activities simultaneously.	CAD: 94.9%, VD: 93.04%
BEHAVE, KU, New Collective Activity	stagNet (Qi *et al.*, 2020)	Group	Semantic graph + spatiotemporal attention for recognizing group and individual activities.	BEHAVE: 98.4%, KU: 84.51%, NCA: 91.25%

Figure 3 shows a two-panel comparison of accuracy. In panel (a), the accuracy in identification of representative models of each category of deep learning architecture on the NTU RGB+D 60 dataset under two common evaluation settings, Cross-Subject (CS) and Cross-View (CV), is recognized. The highest CS accuracy values of 92% and 95.0% are achieved by GCN-based models (CTR-GC, SL-GCNN) and hybrid models, respectively, including AGC-LSTM. Although competitive in earlier studies, RNN-based methods demonstrate a clear performance ceiling in comparison with graph-based and hybrid counterparts. The optimization of models using six metaheuristic algorithms is compared in panel (b) based on their respective HAR benchmark datasets with the best reported accuracies. GWO–WOA hybrid optimization (Yang *et al.*, 2020) and the Rao-3 algorithm (Ding *et al.*, 2020) achieve the highest accuracies (97.5% and 97.16%, respectively), whereas Sea Horse Optimization (Zhang & Li, 2024), applied to hand gesture recognition, also shows competitive results, underlining its potential applicability to more general HAR scenarios.

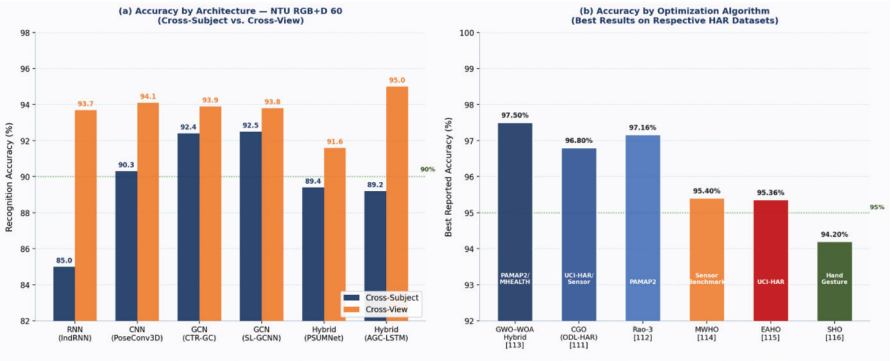


Figure 4: (a) Recognition Accuracy by Deep Learning Architecture on NTU RGB+D 60 (Cross-Subject and Cross-View Protocols); (b) Best Reported Accuracy by Metaheuristic Optimization Algorithm Across HAR Benchmark Datasets

Figure 5 shows how recognition accuracy (NTU RGB+D 60, Cross-Subject protocol) has varied over time in the five largest deep learning architecture categories between 2016 and 2024. All categories have an increasing tendency, which is due to the cumulative effect of architectural innovations, more comprehensive datasets, and better training strategies. The most rapid growth in accuracy is shown by GCN-based methods, which increase their accuracy rates to 92.49% in 2024, with the highest growth from 81.5 to 92.49 between 2018 and 2024, due to the development of dynamic graph learning and channel-wise topology refinement. Transformer-based approaches, which were only introduced later (2020 and later), have demonstrated very fast improvement, bridging the accuracy gap with GCNs. RNN-based approaches, though maturing earlier, exhibit a levelling-off pattern post-2019, in line with structural constraints in their ability to represent spatial joint topology. This direction highlights the increasing significance of graph-structured and attention-based models for 3D HAR and encourages the incorporation of metaheuristic optimization to allow additional accuracy improvements in deploying the systems to the real world.

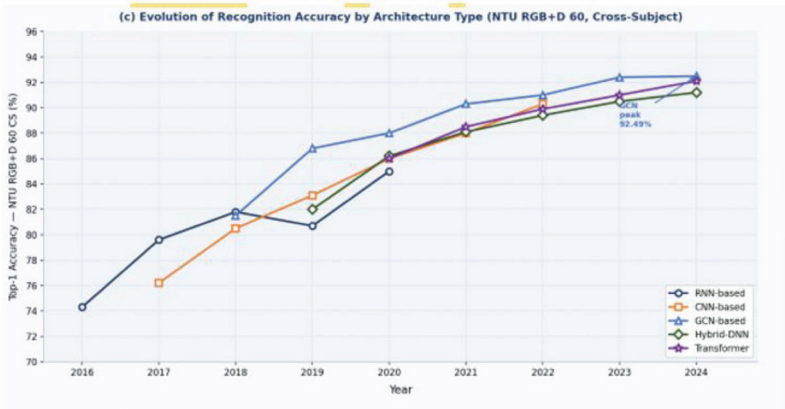


Figure 5: Temporal Evolution of Recognition Accuracy (NTU RGB+D 60, Cross-Subject) across RNN, CNN, GCN, Hybrid-DNN, And Transformer-Based Architectures

9. Conclusion

Recent deep learning-based methods for 3D Human Action Recognition (HAR) were thoroughly reviewed in this study, which also highlighted the strengths and limitations of different models and algorithms. It emphasised how crucial graph-based neural networks, recurrent architectures, and hybrid models are for encapsulating the spatiotemporal dynamics seen in human behaviour. Furthermore, the incorporation of metaheuristic optimisation algorithms, such as Grey Wolf Optimiser, Whale Optimisation, and Chaos Game Optimization, was demonstrated to greatly improve model performance by adjusting hyperparameters and addressing problems such as overfitting and computational inefficiency.

Notwithstanding the noteworthy advancements, a number of unresolved issues still exist, such as the need for more effective and flexible optimisation techniques, explainability of deep learning models, generalisation across various datasets, and real-time processing restrictions. Emerging optimisers such as Sea Horse Optimisation (SHO) show promise, though they require more work and domain-specific modification. Future studies should focus on creating robust and lightweight optimisation algorithms that are suited to the particular complexity of HAR. They should also investigate multi-modal data fusion as a way to enhance cross-dataset generalisation. For real-world applications that require confidence and transparency, improving model interpretability will also be essential. All things considered, improving deep learning architectures and optimisation techniques has enormous potential for creating precise, effective, and deployable HAR systems, opening the door to better surveillance, healthcare, and human-computer interface applications.

Although comprehensive, the survey has some limitations, such as mostly covering supervised deep learning approaches to 3D skeletal human action recognition, and little to no attention to semi-supervised and unsupervised approaches. The testing of approaches is based on benchmark results that have been reported, and this might not be a true representation of real-world environments owing to the presence of sensor noise and data problems. Moreover, the inconsistency between the performance of metaheuristic optimisation algorithms in different architectures hinders performance comparisons. Also, there are no detailed developments in large-scale and foundation models. These gaps must be filled in future surveys through the incorporation of standardized assessment procedures and a focus on data-efficient learning paradigms.

Acknowledgement

The authors declare that there are no acknowledgements for this study.

Authorship Contribution

All authors contributed substantially to the conception, literature review, analysis, manuscript drafting, and final approval of the submitted version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Ethical Approval

This article is a survey/review study and does not involve any human participants, animals, or identifiable personal data. Therefore, ethical approval was not required.

Funding

The authors declare that no specific funding was received for this research work from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

The authors declare that the manuscript is original, has not been published previously, and is not under consideration for publication elsewhere. All sources of information used in this study have been appropriately acknowledged and cited.

Data Availability Statement

No new datasets were generated or analyzed during the current study. All information presented in this survey has been collected from publicly available research articles, datasets, and online resources cited in the references section.

Figure Permission

All figures included in this manuscript are either created by the authors or adapted/redrawn from existing literature with appropriate citation and academic fair-use consideration. The authors are responsible for obtaining any necessary permissions, if required by the publisher.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

AI Disclosure Statement

During the preparation of this manuscript, the author(s) used AI-assisted language tools for grammar correction, language editing, readability improvement, and refinement

of academic writing. After using these tools, the author(s) thoroughly reviewed, verified, and revised all AI-assisted content to ensure accuracy, originality, and scholarly integrity. The author(s) take full responsibility for the integrity and final content of the published article.

References

Ahmad, A. Y. A. B., Alzubi, J., James, S., Nyangaresi, V. O., Kutralakani, C., & Krishnan, A. (2024). Enhancing human action recognition with adaptive hybrid deep attentive networks and Archerfish optimization. *Computers, Materials & Continua, 80*(3). https://doi.org/10.32604/cmc.2024.052771

Ahmad, T., Jin, L., Zhang, X., Lai, S., Tang, G., & Lin, L. (2021). Graph convolutional neural network for human action recognition: A comprehensive survey. *IEEE Transactions on Artificial Intelligence, 2*(2), 128–145. https://doi.org/10.1109/TAI.2021.3076974

Al-Wesabi, F. N., Albraikan, A. A., Hilal, A. M., Al-Shargabi, A. A., Alhazbi, S., Al Duhayyim, M., *et al.* (2021). Design of optimal deep learning based human activity recognition on sensor enabled internet of things environment. *IEEE Access, 9*, 143988–143996. https://doi.org/10.1109/ACCESS.2021.3112973

Amirthalingam, P., Alatawi, Y., Chellamani, N., Shanmuganathan, M., Ali, M. A. S., Alqifari, S. F., *et al.* (2024). Sea horse optimization–deep neural network: A medication adherence monitoring system based on hand gesture recognition. *Sensors, 24*(16), 5224. https://doi.org/10.3390/s24165224

Anbazhagan, K., Swamy, G., Janani, R., & Farakte, A. (2024). Deep learning based human activity recognition in smart home. In *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICDECS59733.2023.10502527

Arshad, M. H., Bilal, M., & Gani, A. (2022). Human activity recognition: Review, taxonomy and open challenges. *Sensors, 22*(17), 6463. https://doi.org/10.3390/s22176463

Basak, H., Kundu, R., Singh, P. K., Ijaz, M. F., Woźniak, M., & Sarkar, R. (2022). A union of deep learning and swarm-based optimization for 3D human action recognition. *Scientific Reports, 12*(1), 5494. https://doi.org/10.1038/s41598-022-09293-8

Bavil, A. F., Damirchi, H., & Taghirad, H. D. (2023). Action capsules: Human skeleton action recognition. *Computer Vision and Image Understanding, 233*, 103722. https://doi.org/10.1016/j.cviu.2023.103722

Beddiar, D. R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: A survey. *Multimedia Tools and Applications, 79*(41), 30509–30555. https://doi.org/10.1007/s11042-020-09004-3

Bian, S., Liu, M., Zhou, B., Lukowicz, P., & Magno, M. (2024). Body-area capacitive or electric field sensing for human activity recognition and human-computer interaction: A comprehensive survey. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8*(1), 1–49. https://doi.org/10.1145/3643555

Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys, 46*(3), 1–33. https://doi.org/10.1145/2499621

Bulugu, I. (2024). Adaptive shift graph convolutional neural network for hand gesture recognition based on 3D skeletal similarity. *Signal, Image and Video Processing, 18*(11), 7583–7595. https://doi.org/10.1007/s11760-024-03412-w

Caetano, C., Brémond, F., & Schwartz, W. R. (2019). Skeleton image representation for 3D action recognition based on tree structure and reference joints. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (pp. 16–23). IEEE. https://doi.org/10.1109/SIBGRAPI.2019.00011

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7291–7299). https://doi.org/10.48550/arXiv.1611.08050

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Challa, S. K., Kumar, A., Semwal, V. B., & Dua, N. (2023). An optimized deep learning model for human activity recognition using inertial measurement units. *Expert Systems, 40*(10), e13457. https://doi.org/10.1111/exsy.13457

Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding, 117*(6), 633–659. https://doi.org/10.1016/j.cviu.2013.01.013

Chen, D., Chen, M., Wu, P., Wu, M., & Zhang, T. (2025). Two-stream spatio-temporal GCN-transformer

networks for skeleton-based action recognition. *Scientific Reports, 15*(1), 87752. https://doi.org/10.1038/s41598-025-87752-8

Chen, J., Yang, W., Liu, C., & Yao, L. (2021). A data augmentation method for skeleton-based action recognition with relative features. *Applied Sciences, 11*(23), 11481. https://doi.org/10.3390/app112311481

Chen, L., Wei, H., & Ferryman, J. (2013). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters, 34*(15), 1995–2006. https://doi.org/10.1016/j.patrec.2013.02.006

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13359–13368). https://doi.org/10.48550/arXiv.2107.12213

Chen, Z., Zhang, L., Jiang, C., Cao, Z., & Cui, W. (2019). WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Transactions on Mobile Computing, 18*(11), 2714–2724. https://doi.org/10.1109/TMC.2018.2878233

Cherla, S., Kulkarni, K., Kale, A., & Ramasubramanian, V. (2008). Towards fast, view-invariant human action recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–8). IEEE. https://doi.org/10.1109/CVPRW.2008.4563179

Cho, S., Maqbool, M., Liu, F., & Foroosh, H. (2020). Self-attention network for skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 635–644). https://doi.org/10.48550/arXiv.1912.08435

Coronado, E., Itadera, S., & Ramirez-Alpizar, I. G. (2023). Integrating virtual, mixed, and augmented reality to human–robot interaction applications using game engines: A brief review of accessible software tools and frameworks. *Applied Sciences, 13*(3), 1292. https://doi.org/10.3390/app13031292

D'Sa, A. G., & Prasad, B. G. (2019). A survey on vision based activity recognition, its applications and challenges. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* (pp. 1–8). IEEE. https://doi.org/10.1109/ICACCP.2019.8882896

Daga, Y., & Meena, S. (2022). Applications of human activity recognition in different fields: A review. In *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (pp. 1–6). IEEE. https://doi.org/10.1109/UPCON56432.2022.9986408

Dhiman, C., & Vishwakarma, D. K. (2019). A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence, 77*, 21–45. https://doi.org/10.1016/j.engappai.2018.08.014

Ding, X., Yang, K., & Chen, W. (2019). An attention-enhanced recurrent graph convolutional network for skeleton-based action recognition. In *Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning* (pp. 79–84). https://doi.org/10.1145/3372806.3372814

Ding, X., Yang, K., & Chen, W. (2020). A semantics-guided graph convolutional network for skeleton-based action recognition. In *Proceedings of the 2020 4th International Conference on Innovation in Artificial Intelligence* (pp. 130–136). https://doi.org/10.1145/3390557.3394129

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al.* (2020). *An image is worth 16×16 words: Transformers for image recognition at scale*. arXiv. https://doi.org/10.48550/arXiv.2010.11929

Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2969–2978). https://github.com/kennymckormick/pyskl

Edel, M., & Köppe, E. (2016). Binarized-BLSTM-RNN based human activity recognition. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 1–7). IEEE. https://doi.org/10.1109/IPIN.2016.7743581

El-Assal, M., Tirilly, P., & Bilasco, I. M. (2024). S3TC: Spiking separated spatial and temporal convolutions with unsupervised STDP-based learning for action recognition. In *International Conference on Pattern Recognition* (pp. 299–314). Springer. https://doi.org/10.1007/978-3-031-78395-1_20

Gao, J., He, T., Zhou, X., & Ge, S. (2019). Focusing and diffusion: Bidirectional attentive graph convolutional networks for skeleton-based action recognition. *arXiv*. https://doi.org/10.48550/arXiv.1912.11521

Gao, X., Jin, Y., Dou, Q., Fu, C.-W., & Heng, P.-A. (2021). Accurate grid keypoint learning for efficient video prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 5908–5915). IEEE. https://doi.org/10.1109/IROS51168.2021.9636874

Gaur, D., & Kumar Dubey, S. (2022). Development of activity recognition model using LSTM-RNN deep learning algorithm. *Journal of Information and Organizational Sciences, 46*(2), 277–291. https://doi.org/10.31341/jios.46.2.1

Gavrilyuk, K., Sanford, R., Javan, M., & Snoek, C. G. M. (2020). Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 839–848). https://doi.org/10.48550/arXiv.2003.12737

Ghadiyaram, D., Tran, D., & Mahajan, D. (2019). Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12046–12055). https://doi.org/10.1109/CVPR.2019.01232

Gilroy, S., Glavin, M., Jones, E., & Mullins, D. (2021). Pedestrian occlusion level classification using keypoint detection and 2D body surface area estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 3833–3839). https://doi.org/10.1109/ICCVW54120.2021.00427

Guha, R., Khan, A. H., Singh, P. K., Sarkar, R., & Bhattacharjee, D. (2021). CGA: A new feature selection model for visual human action recognition. *Neural Computing and Applications, 33*, 5267–5286. https://doi.org/10.1007/s00521-020-05297-5

Gupta, S. (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. *International Journal of Information Management Data Insights, 1*(2), 100046. https://doi.org/10.1016/j.jjimei.2021.100046

Ha, S., Yun, J.-M., & Choi, S. (2015). Multi-modal convolutional neural networks for activity recognition. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 3017–3022). IEEE. https://doi.org/10.1109/SMC.2015.525

Hameed, S., Qolomany, B., Belhaouari, S. B., Abdallah, M., Qadir, J., & Al-Fuqaha, A. (2025). Large language model enhanced particle swarm optimization for hyperparameter tuning for deep learning models. *IEEE Open Journal of the Computer Society, 6*, 574–585. https://doi.org/10.1109/OJCS.2025.3564493

He, J., Zhang, C., He, X., & Dong, R. (2020). Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features. *Neurocomputing, 390*, 248–259. https://doi.org/10.1016/j.neucom.2019.07.103

Hoang, V.-H., Lee, J. W., Piran, M. J., & Park, C.-S. (2023). Advances in skeleton-based fall detection in RGB videos: From handcrafted to deep learning approaches. *IEEE Access, 11*, 92322–92352. https://doi.org/10.1109/ACCESS.2023.3307138

Hossain, H. M. S., Al Haiz Khan, M. D. A., & Roy, N. (2018). DeActive: Scaling activity recognition with active deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(2), 1–23. https://doi.org/10.1145/3214269

Hu, J.-F., Zheng, W.-S., Lai, J., & Zhang, J. (2015). Jointly learning heterogeneous features for RGB-D activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5344–5352). https://doi.org/10.1109/CVPR.2015.7299172

Huang, X., Mei, G., & Zhang, J. (2023). Cross-source point cloud registration: Challenges, progress and prospects. *Neurocomputing, 548*, 126383. https://doi.org/10.1016/j.neucom.2023.126383

Huang, X., Zhou, H., Wang, J., Feng, H., Han, J., Ding, E., *et al.* (2023). Graph contrastive learning for skeleton-based action recognition. *arXiv*. https://doi.org/10.48550/arXiv.2301.10900

Huynh-The, T., Hua, C.-H., & Kim, D.-S. (2019). Encoding pose features to images with data augmentation for 3-D action recognition. *IEEE Transactions on Industrial Informatics, 16*(5), 3100–3111. https://doi.org/10.1109/TII.2019.2910876

Ibh, M., Grasshof, S., Witzner, D., & Madeleine, P. (2023). TemPose: A new skeleton-based transformer model designed for fine-grained motion recognition in badminton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 5199–5208). https://doi.org/10.1109/CVPRW59228.2023.00548

Inoue, M., Inoue, S., & Nishida, T. (2018). Deep recurrent neural network for mobile human activity recognition with high throughput. *Artificial Life and Robotics, 23*, 173–185. https://doi.org/10.1007/s10015-017-0422-x

Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2022). Human activity recognition using tools of convolutional neural networks: A state-of-the-art review, data sets, challenges, and future prospects. *Computers in Biology and Medicine, 149*, 106060. https://doi.org/10.1016/j.compbiomed.2022.106060

Jain, P., Kulis, B., & Grauman, K. (2008). Fast image search for learned metrics. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). IEEE. https://doi.org/10.1109/CVPR.2008.4587841

Jaiswal, S., & Jaiswal, T. (2020). Remarkable skeleton based human action recognition. *Artificial Intelligence Evolution*, 108–121. https://doi.org/10.37256/aie.122020562

Jegham, I., Khalifa, A. B., Alouani, I., & Mahjoub, M. A. (2020). Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation, 32*, 200901. https://doi.org/10.1016/j.fsidi.2019.200901

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., & Black, M. J. (2013). Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3192–3199). https://doi.org/10.1109/ICCV.2013.396

Kahatapitiya, K., & Ryoo, M. S. (2021). Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8385–8394). https://doi.org/10.1109/CVPR46437.2021.00828

Karthickkumar, S., & Kumar, K. (2020). A survey on deep learning techniques for human action recognition. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCI48352.2020.9104135

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., *et al.* (2017). *The Kinetics human action video dataset*. arXiv. https://doi.org/10.48550/arXiv.1705.06950

Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., *et al.* (2024). Human action recognition using fusion of multiview and deep features: An application to video surveillance. *Multimedia Tools and Applications, 83*(5), 14885–14911. https://doi.org/10.1007/s11042-020-08806-9

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys, 54*(10s), 1–41. https://doi.org/10.1145/3505244

Kishore, P. V. V., Kumar, D. A., Sastry, A. S. C. S., & Kumar, E. K. (2018). Motionlets matching with adaptive kernels for 3-D Indian sign language recognition. *IEEE Sensors Journal, 18*(8), 3327–3337. https://doi.org/10.1109/JSEN.2018.2810449

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision* (pp. 2556–2563). IEEE. https://doi.org/10.1109/ICCV.2011.6126543

Kumar, M. T. K., Kishore, P. V. V., Madhav, B. T. P., Kumar, D. A., Kala, N. S., Rao, K. P. K., *et al.* (2021). Can skeletal joint positional ordering influence action recognition on spectrally graded CNNs: A perspective on achieving joint order independent learning. *IEEE Access, 9*, 139611–139626. https://doi.org/10.1109/ACCESS.2021.3119455

Lale, T., Yüksek, G., & Çınar, R. F. (2026). A novel hybrid metaheuristic for optimizing deep CNN hyperparameters to enhance heart disease prediction on a comprehensive merged dataset. *Cluster Computing, 29*(2), 117. https://doi.org/10.1007/s10586-025-05892-y

Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials, 15*(3), 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192

Li, M., & Sun, Q. (2021). 3D skeletal human action recognition using a CNN fusion model. *Mathematical Problems in Engineering, 2021*(1), 6650632. https://doi.org/10.1155/2021/6650632

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2019). Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3595–3603). https://doi.org/10.48550/arXiv.1904.12659

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2021). Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(6), 3316–3333. https://doi.org/10.1109/TPAMI.2021.3053765

Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., *et al.* (2021). GroupFormer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13668–13677). https://doi.org/10.48550/arXiv.2108.12630

Li, S., Li, W., Cook, C., & Gao, Y. (2019). Deep independently recurrent neural network (IndRNN). *arXiv*. https://doi.org/10.48550/arXiv.1910.06251

Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., *et al.* (2023). Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18278–18289). https://doi.org/10.48550/arXiv.2303.00748

Li, Y., Wang, Z., Wang, L., & Wu, G. (2018). Actions as moving points. In *European Conference on Computer Vision* (pp. 68–84). Springer. https://doi.org/10.1007/978-3-030-01216-8_5

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (pp. 2–11). ACM. https://doi.org/10.1145/882082.882086

Liu, M., & Yuan, J. (2018). Recognizing human actions as the evolution of pose estimation maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1159–1168). https://doi.org/10.1109/CVPR.2018.00127

Liu, Y., Yang, J., Perera, M., Ji, P., Kim, D., Xu, M., *et al.* (2022). Representation-centric survey of skeletal action recognition and the ANUBIS benchmark. *arXiv*. https://doi.org/10.2139/ssrn.6465233

Mei, G., Poiesi, F., Saltori, C., Zhang, J., Ricci, E., & Sebe, N. (2023). Overlap-guided Gaussian mixture models for point cloud registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 4511–4520). https://doi.org/10.48550/arXiv.2210.09836

Meng, Z., Zhang, M., Guo, C., Fan, Q., Zhang, H., Gao, N., *et al.* (2020). Recent progress in sensing and computing techniques for human activity recognition and motion analysis. *Electronics, 9*(9), 1357. https://doi.org/10.3390/electronics9091357

Nadeem, A., Jalal, A., & Kim, K. (2021). Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy Markov model. *Multimedia Tools and Applications, 80*(14), 21465–21498. https://doi.org/10.1007/s11042-021-10687-5

Nadeem, M. S. (2024). *Hybrid architecture for human action recognition using skeleton data* [Master's thesis]. https://hdl.handle.net/10155/1901

Nguyen, H.-C., Nguyen, T.-H., Scherer, R., & Le, V.-H. (2023). Deep learning for human activity recognition on 3D human skeleton: Survey and comparative study. *Sensors, 23*(11), 5121. https://doi.org/10.3390/s23115121

Ozcan, T., & Basturk, A. (2020). Human action recognition with deep learning and structural optimization using a hybrid heuristic algorithm. *Cluster Computing, 23*(4), 2847–2860. https://doi.org/10.1007/s10586-020-03050-0

Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artificial Intelligence Review, 54*(3), 2259–2322. https://doi.org/10.1007/s10462-020-09904-8

Paul, M., Haque, S. M. E., & Chakraborty, S. (2013). Human detection in surveillance videos and its applications—A review. *EURASIP Journal on Advances in Signal Processing, 2013*(1), 176. https://doi.org/10.1186/1687-6180-2013-176

Plizzari, C., Cannici, M., & Matteucci, M. (2021). Spatial temporal transformer network for skeleton-based action recognition. In *International Conference on Pattern Recognition* (pp. 694–701). Springer. https://doi.org/10.1007/978-3-030-68796-0_50

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing, 28*(6), 976–990. https://doi.org/10.1016/j.imavis.2009.11.014

Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., & Van Gool, L. (2020). StagNet: An attentive semantic RNN for group activity and individual action recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(2), 549–565. https://doi.org/10.1109/TCSVT.2019.2894161

Ramasamy Ramamurthy, S., & Roy, N. (2018). Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1254. https://doi.org/10.1002/widm.1254

Raziani, S., & Azimbagirad, M. (2022). Deep CNN hyperparameter optimization algorithms for sensor-based human activity recognition. *Neuroscience Informatics, 2*(3), 100078. https://doi.org/10.1016/j.neuri.2022.100078

Ren, B., Liu, M., Ding, R., & Liu, H. (2024). A survey on 3D skeleton-based action recognition using learning methods. *Cyborg and Bionic Systems, 5*, 100. https://doi.org/10.34133/cbsystems.0100

Ren, B., Liu, Y., Song, Y., Bi, W., Cucchiara, R., Sebe, N., *et al.* (2023). Masked jigsaw puzzle: A versatile position embedding for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20382–20391). https://doi.org/10.48550/arXiv.2205.12551

Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., & Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 479–485). https://doi.org/10.1109/CVPRW.2013.77

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1010–1019). https://doi.org/10.48550/arXiv.1604.02808

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12026–12035). https://doi.org/10.48550/arXiv.1805.07694

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision*. https://link.springer.com/conference/accv

Shreyas, D. G., Raksha, S., & Prasad, B. G. (2020). Implementation of an anomalous human activity recognition system. *SN Computer Science, 1*(3), 168. https://doi.org/10.1007/s42979-020-00169-0

Shu, X., Zhang, L., Sun, Y., & Tang, J. (2021). Host–parasite: Graph LSTM-in-LSTM for group activity recognition. *IEEE Transactions on Neural Networks and Learning Systems, 32*(2), 663–674. https://doi.org/10.1109/TNNLS.2020.2978942

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1227–1236).

Sighencea, B. I., Stanciu, I. R., & Căleanu, C. D. (2023). D-STGCN: Dynamic pedestrian trajectory prediction using spatio-temporal graph convolutional networks. *Electronics, 12*(3), 611. https://doi.org/10.3390/electronics12030611

Singh, R., Sonawane, A., & Srivastava, R. (2020). Recent evolution of modern datasets for human activity recognition: A deep survey. *Multimedia Systems, 26*(2), 83–106. https://doi.org/10.1007/s00530-019-00635-7

Subetha, T., & Chitrakala, S. (2016). A survey on human activity recognition from videos. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICICES.2016.7518920

Sundaresan, K., Jayarajan, P., & Nallakumar, R. (2025). *Exploring human activity recognition systems: Insights from computer vision approaches*. Research Square. https://doi.org/10.21203/rs.3.rs-5923670/v1

Tahir, B. S., Ageed, Z. S., Hasan, S. S., & Zeebaree, S. R. M. (2023). Modified wild horse optimization with deep learning enabled symmetric human activity recognition model. *Computers, Materials & Continua, 75*(2), 4009–4024. https://doi.org/10.32604/cmc.2023.037433

Tien, P. W., Wei, S., Calautit, J. K., Darkwa, J., & Wood, C. (2021). Vision-based human activity recognition for reducing building energy demand. *Building Services Engineering Research and Technology, 42*(6), 691–713. https://doi.org/10.1177/01436244211026120

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 10347–10357). PMLR. https://proceedings.mlr.press/v139/touvron21a.html

Trivedi, N., & Sarvadevabhatla, R. K. (2022). Psumnet: Unified modality part streams are all you need for efficient pose-based action recognition. In *European Conference on Computer Vision* (pp. 211–227). Springer. https://github.com/skelemoa/psumnet

Vaswani, A. (2017). *Attention is all you need*. arXiv. https://cir.nii.ac.jp/crid/1370580229800306054

Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications, 25*(2), 743–755. https://doi.org/10.1007/s11036-019-01445-x

Wang, B., Huang, L., & Hoai, M. (2020). Active vision for early recognition of human actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1081–1091). https://doi.org/10.1109/CVPR42600.2020.00116

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision, 103*(1), 60–79. https://doi.org/10.1007/s11263-012-0594-8

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1290–1297). IEEE. https://doi.org/10.1109/CVPR.2012.6247813

Wang, J., Nie, X., Xia, Y., Wu, Y., & Zhu, S.-C. (2014). Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2649–2656). https://doi.org/10.1109/CVPR.2014.339

Wang, L., Ding, Z., Tao, Z., Liu, Y., & Fu, Y. (2019). Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6212–6221). https://doi.org/10.1109/ICCV.2019.00631

Wang, W., Mei, G., Ren, B., Huang, X., Poiesi, F., Van Gool, L., *et al.* (2023). *Zero-shot point cloud registration*. arXiv. https://doi.org/10.48550/arXiv.2312.06063

Wu, C., Wu, X.-J., & Kittler, J. (2019). Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Wu, J., Wang, L., Wang, L., Guo, J., & Wu, G. (2019). Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9964–9974). https://doi.org/10.48550/arXiv.1904.10117

Xia, L., Chen, C.-C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 20–27). IEEE. https://doi.org/10.1109/CVPRW.2012.6239233

Xiang, W., Li, C., Zhou, Y., Wang, B., & Zhang, L. (2023). Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10276–10285). https://doi.org/10.48550/arXiv.2208.05318

Xing, Y., & Zhu, J. (2021). Deep learning-based action recognition with 3D skeleton: A survey. *CAAI Transactions on Intelligence Technology, 6*, 80–92. https://doi.org/10.1049/cit2.12014

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v32i1.12328

Yang, D., Li, M. M., Fu, H., Fan, J., Zhang, Z., & Leung, H. (2020). *Unifying graph embedding features with graph convolutional networks for skeleton-based action recognition*. arXiv. https://doi.org/10.48550/arXiv.2003.03007

Yang, X., Li, S., Niu, S., & Yue, X. (2025). Graph network learning for human skeleton modeling: A survey. *Artificial Intelligence Review, 59*(1), 31. https://doi.org/10.1007/s10462-025-11442-0

Yang, Z., Li, Y., Yang, J., & Luo, J. (2018). Action recognition with spatio–temporal visual attention on skeleton image sequences. *IEEE Transactions on Circuits and Systems for Video Technology, 29*(8), 2405–2415. https://doi.org/10.1109/TCSVT.2018.2864148

Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., & Tang, H. (2020). Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 55–63). ACM. https://doi.org/10.1145/3394171.3413941

Ye, L., Rochan, M., Liu, Z., & Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10502–10511). https://doi.org/10.1109/CVPR.2019.01075

Yu, S., Xie, L., Liu, L., & Xia, D. (2019). Learning long-term temporal features with deep neural networks for human action recognition. *IEEE Access, 8*, 1840–1850. https://doi.org/10.1109/ACCESS.2019.2962284

Yuan, L., He, Z., Wang, Q., Xu, L., & Ma, X. (2022). Spatial transformer network with transfer learning for small-scale fine-grained skeleton-based Tai Chi action recognition. In *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society* (pp. 1–6). IEEE. https://doi.org/10.1109/IECON49645.2022.9968668

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., & Samaras, D. (2012). Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 28–35). IEEE. https://doi.org/10.1109/CVPRW.2012.6239234

Zahin, A., Tan, L. T., & Hu, R. Q. (2019). Sensor-based human activity recognition for smart healthcare: A semi-supervised machine learning approach. In *Artificial intelligence for communications and networks* (pp. 450–472). Springer. https://doi.org/10.1007/978-3-642-29336-8_15

Zappardino, F., Uricchio, T., Seidenari, L., & Del Bimbo, A. (2021). Learning group activities from skeletons without individual action labels. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 10412–10417). IEEE. https://doi.org/10.1109/ICPR48806.2021.9413195

Zhang, J., & Li, Y. (2024). SL-GCNN: A graph convolutional neural network for granular human motion recognition. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3514082

Zhang, J., Jia, Y., Xie, W., & Tu, Z. (2022). Zoom transformer for skeleton-based group activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(12), 8646–8659. https://doi.org/10.1109/TCSVT.2022.3193574

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View adaptive recurrent neural networks for high performance human action rec o g n i t i o n from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2117–2126). https://doi.org/10.48550/arXiv.1703.08274

Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., & Zheng, N. (2020). Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1112–1121). https://doi.org/10.48550/arXiv.1904.01189

Zhang, P., Xue, J., Lan, C., Zeng, W., Gao, Z., & Zheng, N. (2019). EleAtt-RNN: Adding attentiveness to neurons in recurrent neural networks. *IEEE Transactions on Image Processing, 29*, 1061–1073. https://doi.org/10.1109/TIP.2019.2937724

Zhang, W., Liu, Z., Zhou, L., Leung, H., & Chan, A. B. (2017). Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. *Image and Vision Computing, 61*, 22–39. https://doi.org/10.1016/j.imavis.2017.02.002

Zhao, R., Wang, K., Su, H., & Ji, Q. (2019). Bayesian graph convolution LSTM for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6881–6890). https://doi.org/10.1109/ICCV.2019.00698

Zhou, X., Liang, W., Kevin, I., Wang, K., Wang, H., Yang, L. T., *et al.* (2020). Deep-learning-enhanced human activity recognition for internet of healthcare things. *IEEE Internet of Things Journal, 7*(7), 6429–6438. https://doi.org/10.1109/JIOT.2020.2985082

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., & Wang, Y. (2023). MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15085–15099). https://doi.org/10.48550/arXiv.2210.06551