THE SCIENTIFIC TEMPER (2025) Vol. 16 (12): 5241-5257

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.12

https://scientifictemper.com/

RESEARCH ARTICLE

Human Activity Recognition through Skeleton-Based Motion Analysis Using YOLOv8 and Graph Convolutional Networks

Subna M P[1*], Kamalraj N[2]

Abstract

Human Activity Recognition has become an important research domain in developing intelligent systems for sectors such as healthcare, behavioral analytics, and surveillance monitoring. Traditional vision-based HAR approaches have limitations in terms of subject variability, occlusion, and background clutter. To address this, a novel skeleton-based motion analysis model is proposed to enhance the precision and temporal understanding of human motions by combining real-time keypoint extraction with graph-structured spatial-temporal learning. The proposed YOLOv8 + Graph Temporal Convolution for Human Activity Recognition (YGTC-HAR) consists of four essential stages, including: (1) YOLOv8-Pose to detect human figures in real-time, and (2) Graph Convolutional Network (GCN) is used to transform the joint coordinates into a graph representation graph representation. (3) The Temporal Convolutional Network (TCN) is designed to learn the sequential motion dynamics and time-dependent characteristics of human activities. Additionally, Genetic Algorithm (GA) and Bayesian Optimization (BO) are adopted to fine-tune hyperparameters, including learning rate, dropout ratio, and convolutional filters. MHealth and WISDM datasets are utilized in this research to enable comprehensive testing across static and dynamic movements. The proposed YGTC-HAR is implemented using Python (with TensorFlow and PyTorch) for deep learning, and MATLAB R2023b is used for signal processing, graphical visualization, and performance validation. The proposed work is compared against existing HLA, SMO-DNN, AMC-CNN, and YOLOv8-ViT models. The model achieves 97.6% accuracy, 98.4% sensitivity, 97.8% specificity, 97.2% F1-score, 96.4% MCC, and an AUC of 0.96, which outperforms the existing models by over 4.3%. The proposed YGTC-HAR serves as a single end-to-end HAR framework that delivers superior generalization, real-time performance, and reliability for HCIA (Human-Centered Intelligent Applications). The novelty of the model lies in the combination of YOLOv8-driven skeleton extraction, GCN-based spatial modeling, TCN-driven temporal learning, and adaptive optimization.

Keywords: Human Activity Recognition, Deep Learning, Graph Convolutional Networks, Skeleton-based Analysis, Temporal Convolutional Networks, YOLOv8.

Introduction

In recent years, HAR has emerged as a transformative research area with a broad spectrum across healthcare, assisted living, and intelligent systems. Real-time motion capturing and activity detection models exhibit improved performance with the aid of inertial sensor data and raw video. The majority of conventional HAR techniques have limitations related to occlusion, viewpoint dependency, illumination challenges, and inconsistent motion representation, which reduce the reliability of HAR, especially in dynamic real-world environments where human movement is irregular and complex. To overcome the limitations, the YTGC-HAR model is proposed by leveraging the merits of deep learning-based skeleton motion analysis to learn about the human body and abstracting it into a set of joints & connections, while eliminating background noise and preserving spatial

[1]Research Scholar (Full Time), Department of Computer Science, Park's College (Autonomous), Chinnakkarai, Tirupur, Tamilnadu, India - 641605

[2]Vice Principal, Park's College (Autonomous), Chinnakkarai, Tirupur, Tamilnadu, India - 641605

***Corresponding Author:** Subna M P, Research Scholar (Full Time), Department of Computer Science, Park's College (Autonomous), Chinnakkarai, Tirupur, Tamilnadu, India - 641605, E-Mail: rehnasubi@gmail.com

How to cite this article: Subna. M.P., Kamalraj, N. (2025). Human Activity Recognition through Skeleton-Based Motion Analysis Using YOLOv8 and Graph Convolutional Networks. The Scientific Temper, **16**(12):5241-5257.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.12

Source of support: Nil

Conflict of interest: None.

Received: 13/11/2025 **Accepted:** 14/12/2025 **Published:** 25/12/2025

and temporal relationships. YOLOv8, GCN, TCN, GA, and BO combinational techniques are employed in the proposed work to address the growing demand for intelligent human-centric automation. The YTGC-HAR system thus represents a significant step toward reliable, efficient, and scalable HAR suitable for embedded and real-world deployments, such as surveillance, assisted living, rehabilitation tracking, elderly care, and remote health assessment, ensuring proactive healthcare decision-making and safety monitoring.

Recent developments in HAR demonstrates notable progress through various deep learning architectures and optimized feature extraction techniques. An attention-driven deep learning model with temporal and spatial features significantly enhanced feature discrimination and attains higher precision in sensor-based HAR. The use of dual-attention layers improved focus on critical time segments; however, the method's static weighting limits its adaptability under irregular activity transitions and real-world healthcare variations (Akter et al., 2023). A hybrid learning algorithm combining convolutional and recurrent structures (HLA) provided an effective model for both spatial and temporal features in wearable sensor data. This approach captured continuous movement patterns accurately but exhibited reduced generalization across subjects, primarily due to noise sensitivity and sensor orientation inconsistencies (Athota & Sumathi, 2022). A multichannel convolutional neural network enhanced with extensive data augmentation improved the recognition of overlapping activities by capturing feature diversity across multiple sensor streams. Although the approach yielded higher accuracy, its heavy computational cost limited its application in low-power wearable healthcare systems (Shi et al., 2022).

A deep learning technique incorporating Spider Monkey Optimization (SMO-DNN) provided efficient feature selection and convergence control. The algorithm effectively improved accuracy on benchmark datasets; however, it required extensive parameter tuning and longer training times, which restricted its use for large-scale or real-time activity recognition (Kolkar & Geetha, 2023). A hybrid deep learning architecture utilizing convolutional layers with logistic gating enhanced information flow and mitigated gradient vanishing in IoT-based HAR. Despite achieving strong performance, the absence of spatial body-joint modeling limited its interpretability in skeleton-based applications (Ding, Abdel-Basset, & Mohamed, 2023). An orientation-invariant deep learning framework employing angular normalization stabilized predictions across devices placed in varying positions. While this method enhanced orientation robustness, it showed weaker adaptability to multimodal sensor combinations with high-dimensional signals (He, Sun, & Zhang, 2024). Graph-based neural representations modeled human joints as interconnected nodes to effectively capture the relationships between body structure and motion. Although this method preserved spatial information, it lacked adequate temporal modeling, which is critical for distinguishing similar dynamic activities such as running and walking (Bsoul, 2025). A hybrid YOLOv8-based deep learning framework further extended HAR accuracy using pose extraction and data augmentation on MHealth and WISDM datasets. It demonstrated robust cross-subject generalization but remained limited by sequential CNN–RNN architectures that inadequately represented spatio-temporal dependencies (Subna & Kamalraj, 2025).

Problem Statement

Existing HAR models face persistent challenges in accurately classifying essential activities such as walking, sitting, lying down, or exercising, which are key indicators of patient mobility and recovery progress in patient monitoring environments. Conventional deep learning models trained on raw image frames and inertia sensor-based scanning often overfit in feature analysis, leading to poor spatial and temporal reasoning with limited interpretability. In the skeleton-based approach, the motion abstraction is clean and straightforward. In contrast, the present graph-based architectures failed to handle spatial and temporal dependencies simultaneously, resulting in incomplete modeling of human kinematics. The fundamental problem addressed in this research is the inadequate integration of spatial and temporal learning for precise activity recognition. YGTC-HAR overcomes the limitations of optimized fusion of detection, representation, and sequence learning, offering real-time adaptability and scalability across diverse datasets. The proposed work will provide a unified framework for extracting, encoding, and robustly interpreting skeletal motion patterns. YGTC-HAR serves as an optimized pipeline for accurate and robust activity recognition, filling the gap left by existing baseline models.

Key Objectives of the Proposed Model

The primary objective of the YTGC-HAR model is to design and implement a robust, performance-oriented skeleton-based HAR framework that recognizes human activities with greater accuracy by integrating spatial-temporal feature learning with bio-inspired AGA and Bayesian optimization for fine-tuning. The following are the key objectives of the proposed deep learning framework.

- To extract real-time skeletal key points from human frames using YOLOv8-Pose and refine them using Media-Pipe pose estimation.
- To model skeletal motion as a spatial-temporal graph, where nodes represent body joints and edges define anatomical relations.
- To employ a Graph Convolutional Network (GCN) for learning spatial dependencies between joints and a Temporal Convolutional Network (TCN) to analyse

motion transitions and long-range dependencies across frames.

- To optimize hyperparameters through Genetic Algorithm (GA) and Bayesian Optimization (BO) for improved efficiency and convergence.
- To validate performance using MHealth and WISDM datasets under real-time simulation in Python and MATLAB environments.

These combinational steps collectively demonstrate that the suggested YTGC-HAR model achieves high recognition accuracy, faster inference, and robust adaptability across subjects and activities in real-time, setting a novel pattern for human motion understanding in next-generation intelligent systems.

Related Works

Recent work on HAR for industrial and clinical settings showcases the task realism, such as manual material handling, where deep learning techniques must separate subtle load-bearing postures from ordinary motion. A comprehensive pipeline utilizing deep learning enhances the detection of ergonomically risky actions, yet remains sensitive to occlusions and rapid viewpoint shifts common on factory floors, underscoring the need for stronger temporal context and skeleton priors (Bassani et al., 2025). Surveys provide machine-learning models that highlight HAR in terms of feature learning, transferability, and deployment, yet also notable gaps in cross-dataset robustness and interpretability for decision support in healthcare. The consensus calls for hybrid spatio-temporal modeling and clear post-hoc explanations to gain clinician trust (Hossen & Abas, 2025). Security-oriented healthcare studies demonstrate that deep learning can learn distinctive intrinsic patterns from constrained sensors, indicating that well-structured physiological signals are highly discriminative. Still, these systems often lack temporal generalization across sessions and devices, a limitation directly relevant to HAR wearables (Indhumathi et al., 2025). Decision analytics pipelines in agriculture are illustrated, demonstrating how domain-aware features and intelligent fusion enhance reliability under noisy conditions. Despite its high accuracy, many such systems rely on static thresholds and handcrafted rules for final decisions, which limits their adaptability when motion dynamics drift—a pattern also observed in naive HAR post-processing (Jijendra & Nithyanandh, 2025).

Hybrid deep learning for sensor HAR combining convolutional backbones with recurrent or attention heads boosts accuracy and energy efficiency. However, reliance on fixed window sizes reduces sensitivity to variable-speed actions, motivating the use of dilated temporal encoders that flexibly cover multiple time scales (Khan, Afzal, & Lee, 2022). Surveillance-focused interaction recognition benefits from coupling deep features with classical machine-learning classifiers to stabilize small-sample regimes. Nevertheless, without an explicit joint-level structure, models struggle to parse fine-grained interactions, such as handovers or near-collisions—precisely where skeleton graphs are helpful (Khean et al., 2024). Vision research on enhancing human sight perception for machine vision highlights the importance of multi-resolution cues and attention in suppressing background clutter. While object-centric attention improves precision, activity recognition additionally requires modeling dependencies across joints over time, beyond region saliency alone (Krishnaveni et al., 2023). Residual networks with squeeze-and-excitation mechanisms provide channel-wise recalibration, sharpening salient motion features and yielding explainability through activation maps.

Despite this, channel attention alone cannot encode relational kinematics; joint-edge reasoning is needed for nuanced pose transitions (Mekruk & Jitpattanakul, 2025). Wearable biosensor applications (e.g., smart knee bandages) demonstrate that localized sensors can predict rehabilitation activities; however, segment generalization remains brittle when patients alter their gait or cadence. Structured spatio-temporal models can mitigate such drift by anchoring predictions to joint graphs and tempo-robust encoders (Savanich, Jantawong, & Jitpattanakul, 2022). A detailed overview of deep HAR highlights the maturity of CNN/RNN baselines and emphasize the importance of data augmentation for achieving class balance. Still, many systems underperform when activities overlap in space and time, underscoring the need for temporal receptive fields with dilation and residual connections (Moola & Hossain, 2022). Object-detection pipelines with modern detectors, such as YOLOv8, illustrate that robust localization is feasible in real-time, but downstream activity semantics require structured modeling beyond bounding boxes. Bridging fast detection with graph-temporal reasoning is therefore a logical next step (Nithyanandh, 2025). Multimodal biometric security, which combines gait and face recognition, suggests that complementary views can compensate for modality-specific failures. For HAR, analogous fusion (pose + inertial) can stabilize predictions under camera occlusion, though careful alignment and calibration remain open challenges (Nivedita et al., 2025). Wearable-sensor HAR continues to benefit from deep learning with careful segmentation, normalization, and augmentation. However, many pipelines treat each window independently, missing long-range dependencies and inter-joint constraints that distinguish closely related actions (Nouriani, McGovern, & Rajamani, 2022).

Research Gap Analysis

Across these studies, few major strengths include improved feature saliency, augmentation, and domain-specific tailoring. Current HAR models predominantly focus on either spatial or temporal components, often overlooking

Table 1: Analysis of Related Studies Supporting Human Activity Recognition

Authors	*Methods Adopted*	*Merits*	*Limitations*
Surek et al. (2023)	Deep learning-based video HAR using CNN and LSTM models for sequential motion interpretation.	Achieved high frame-wise recognition accuracy; effective for video sequence understanding.	Computationally intensive for real-time deployment; limited to controlled datasets.
Uddin et al. (2024)	Hybrid deep learning combining CNN, Conv-LSTM, and LRCN for temporal-spatial feature extraction.	Enhanced activity recognition accuracy; captured both short-term and long-term dependencies.	Requires large labeled datasets and high GPU resources; prone to overfitting.
Xu et al. (2025)	Attention-enhanced deep neural network integrating context-aware motion weighting for HAR.	Improved interpretability and dynamic attention adaptability in feature extraction.	Performance declines under noisy sensor signals; lacks temporal robustness.
Zhang et al. (2024)	Multi-channel hybrid deep learning for multi-sensor data fusion and robust activity classification.	Superior fusion of heterogeneous sensor modalities; reduced feature redundancy.	Fusion complexity increases with more modalities; needs feature alignment optimization.
Sarveshwaran et al. (2022)	Comprehensive investigation of deep learning architectures (CNN, RNN) for HAR performance evaluation.	Provided baseline analysis of DL performance; highlighted dataset dependency challenges.	Limited to small-scale datasets; lacks graph-based structural analysis.
Selvam & Joy (2024)	Deep learning AEN with Mask R-CNN for multivariable feature selection and region-based detection.	Accurate region-based detection; strong performance in multivariable feature environments.	High memory consumption; requires fine-tuning for varied image resolutions.
Sharen et al. (2024)	WISNet deep neural network for human activity recognition from wearable sensor data.	High accuracy for wearable sensor HAR; robust to sensor noise and drift.	Not tested for cross-subject generalization; lacks temporal modeling.
Omprakash et al. (2023)	Energy-aware adaptive sleep scheduling with improvised Firefly Algorithm for efficient IoT communication.	Extended network lifetime and energy efficiency in IoT environments.	Designed for IoT energy systems for sensor-based target class detection tasks.
Prabhu et al. (2025)	Bio-inspired routing using intelligent algorithms for secure and energy-optimized 6G communication.	Improved routing reliability, authentication, and energy utilization in 6G IoT systems.	Focuses on sensor security; limited applicability to human motion recognition.
Eldho & Nithyanandh (2024)	3D CNN model applied on CT-DICOM dataset for lung cancer detection and severity classification.	Accurate 3D volumetric analysis for clinical diagnostics; reduced false positives.	Deep object detection is not generalized for non-medical datasets or HAR applications.
Eldho et al. (2025)	Quantum Hybrid Harris Hawk Optimization with Graph Neural Network for WSN reliability.	Enhanced fault detection and routing efficiency under varying WSN conditions.	Complexity in quantum optimization; increased computational overhead.
Devi et al. (2024)	GAN-enabled AI-based bio-inspired protocol for efficient and secure IoT data transmission.	Ensured secure and low-latency IoT communication; minimized data loss.	High model complexity; limited generalization under dynamic conditions.
Arularasan et al. (2024)	Deep learning model for sign language recognition using spatial feature extraction and classification.	Improved sign recognition precision for hearing-impaired assistance applications.	Limited dataset diversity; requires multi-lingual gesture expansion.

their joint correlation. Moreover, optimization processes in existing models are computationally demanding and prone to local convergence. Persistent gaps involve (i) insufficient modeling of joint relationships and temporal dependencies together, (ii) limited cross-subject and cross-dataset generalization, and (iii) a lack of real-time, explainable inference suitable for healthcare workflows.

The proposed YTGC-HAR addresses these gaps by fusing fast pose extraction (YOLOv8) with graph-based spatial reasoning (GCN) and dilated temporal encoding (TCN), delivering interpretable, scalable, and robust activity recognition across realistic healthcare scenarios. This unified design ensures precise, real-time, and scalable HAR suitable for healthcare, surveillance, and activity monitoring applications.

Proposed Methodology

The proposed YTGC-HAR (YOLOv8-Temporal Graph Convolutional Human Activity Recognition) framework integrates spatial and temporal learning techniques to classify human activities from skeleton-based motion data accurately. The methodology is designed to extract meaningful motion representations from raw sensor readings and body joint positions, combining the strengths

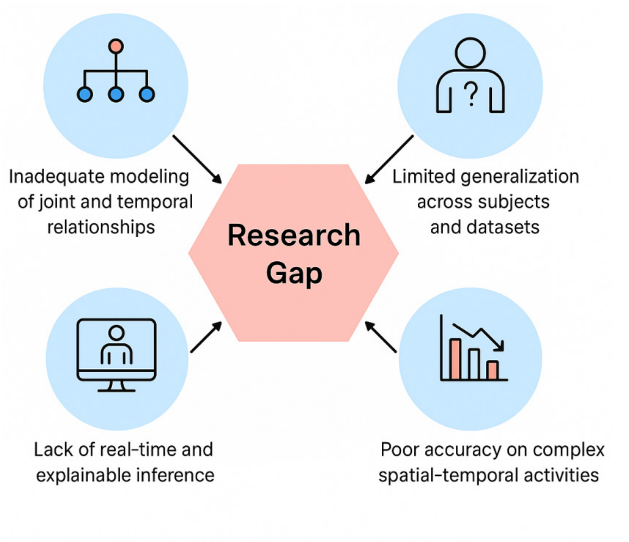

Inadequate modeling of joint and temporal relationships
Limited generalization across subjects and datasets
Research Gap
Lack of real-time and explainable inference
Poor accuracy on complex spatial-temporal activities

Figure 1: Research Gap based on Literature Study

of deep learning and optimization-based tuning for robust human activity recognition. As a first step, data from the MHealth and WISDM datasets are pre-processed and normalized. From the MHealth dataset, attributes such as accelerometer, gyroscope, and magnetometer readings (along X, Y, Z axes) are utilized to represent linear acceleration, angular velocity, and orientation. Similarly, in the WISDM dataset, sensor readings from smartphones and smartwatches are extracted, including acceleration, body movement, and posture variations. These multivariate features form the temporal signal inputs for constructing skeletal joint patterns that represent each user's motion frame by frame. After the pre-processing stage, YOLOv8-Pose is used to detect human figures and extract key skeletal joints, including the head, shoulders, elbows, wrists, knees, and ankles. Each detected joint acts as a node, and anatomical connections between joints act as edges in a graph structure. The GCN then processes this graph and learns the spatial dependencies between connected joints, identifying correlated movement regions. To capture motion steadiness, the TCN analyzes sequential frames, identifying long-term dependencies and distinguishing between similar actions, such as jogging, running, and climbing. The framework is further optimized using an AGA to fine-tune parameters such as the learning rate, dropout, and convolutional depth, ensuring fast convergence and improved generalization. Figure 2 shows the systematic flow diagram of the proposed YTGC-HAR model.

Materials and Methods for Implementation

To evaluate the effectiveness of the proposed YTGC-HAR deep learning model, a robust methodology is adopted by integrating data acquisition, pre-processing, model training, and optimization. The framework combines YOLOv8-based skeleton extraction with GCN–TCN learning to capture spatiotemporal dynamics, ensuring accurate and real-time human activity recognition, validated on benchmark healthcare datasets. This section describes datasets, pre-processing techniques, model architecture, training process, optimization, and evaluation strategies employed to implement the proposed HAR model.

Dataset Description

Two benchmark datasets, MHealth and WISDM, are utilized for training, testing, and validation purposes. The MHealth dataset comprises accelerometer, gyroscope, and magnetometer signals from 10 participants performing 12 activities, whereas WISDM provides over one million motion readings from 36 users, captured through smartphones and smartwatches.

Data Pre-processing

Raw signals are filtered using a Butterworth noise filter, normalized using z-score scaling, and segmented into fixed time windows. Video frames were extracted for skeleton generation using YOLOv8-Pose, followed by keypoint refinement with MediaPipe.

Model Training

The skeleton key-points are converted into a spatio-temporal graph, processed through GCN for spatial learning and TCN for motion sequence modeling.

Optimization

An Adaptive Genetic Algorithm (AGA) bio-inspired optimization was used to fine-tune learning parameters for faster convergence and improved generalization.

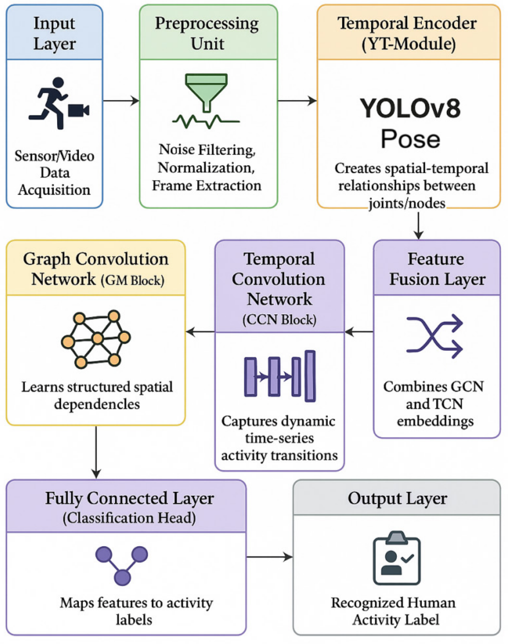

Input Layer
Sensor/Video Data Acquisition
Preprocessing Unit
Noise Filtering, Normalization, Frame Extraction
Temporal Encoder (YT-Module)
YOLOv8 Pose
Creates spatial-temporal relationships between joints/nodes
Graph Convolution Network (GM Block)
Learns structured spatial dependencies
Temporal Convolution Network (CCN Block)
Captures dynamic time-series activity transitions
Feature Fusion Layer
Combines GCN and TCN embeddings
Fully Connected Layer (Classification Head)
Maps features to activity labels
Output Layer
Recognized Human Activity Label

Figure 2: Systematic Flow of Methodology

Implementation Tools

Model training and analysis were conducted in Python (TensorFlow and PyTorch), while visualization and statistical validation were performed in MATLAB R2023b.

Evaluation

Performance is evaluated using six key metrics, such as accuracy, sensitivity, specificity, F1-score, MCC, and AUC, to ensure robustness and real-time efficiency.

Dataset Acquisition and Preprocessing Techniques

The proposed YTGC-HAR framework is built upon two widely recognized datasets, MHealth and WISDM, to ensure robust and generalizable human activity modeling. These datasets were selected due to their extensive coverage of physical movements, sensor diversity, and data consistency, which collectively provide a strong foundation for both sensor-based and vision-based human activity recognition. The MHealth (Mobile Health) dataset was designed for healthcare-oriented activity analysis and physical monitoring applications. It contains time-synchronized recordings from 10 volunteer subjects, each performing 12 distinct physical activities, including standing, sitting, cycling, walking, jogging, running, climbing stairs, lying down, and jumping. Data were captured using accelerometer, gyroscope, and magnetometer sensors placed on the chest, left ankle, and right wrist of each subject. Each sensor stream records signals at a sampling frequency of $50\,Hz$, yielding multivariate time-series data. These signals represent tri-axial motion information $(X,Y,Z\,axes)$ for linear acceleration, angular velocity, and orientation, producing over 23,000 labeled activity segments. MHealth Data Source: https://archive.ics.uci.edu/dataset/319/mhealth+dataset

The WISDM dataset complements MHealth by providing a large-scale motion dataset captured from everyday smartphone and smartwatch sensors. It includes data from 36 individuals performing six fundamental activities: walking, jogging, standing, sitting, ascending stairs, and descending stairs. Each reading contains timestamped accelerometer and gyroscope data at sampling rates of $20-50\,Hz$, totaling over 1 million labeled motion instances. The WISDM dataset offers real-world complexity by accounting for device orientation changes and natural variations in user motion, making it highly valuable for testing model generalization. By integrating both datasets, the YTGC-HAR model benefits from a dual-domain input, such as (1) wearable sensor data for quantitative motion analysis and (2) video frame-based skeletal data for spatial feature extraction. This fusion ensures a comprehensive understanding of human movements across physical, behavioral, and contextual dimensions. Figure 3 shows the pre-processing flow of the proposed model. WISDM Data Source: https://archive.ics.uci.edu/datase t/507/wisdm+smartphon e+and+smar twatch+activity+an d+biom etrics+dataset

To prepare the datasets for model training, five major pre-processing steps are employed to ensure signal quality, normalization, and structural uniformity. These steps mitigate sensor noise, irregular sampling, and dynamic variations among users.

Step 1: Noise Filtering using Butterworth Filter - BWF

The Raw inertial signals are prone to high-frequency noise generated by sensor vibration or hardware inconsistencies. A Butterworth low-pass filter of order $n=4$ with a cutoff frequency $f_c=20\,Hz$ is applied to preserve smooth motion transitions. The transfer function is mathematically represented as,

$$H(s)=\frac{1}{\sqrt{1+\left(\frac{s}{\omega_c}\right)^{2n}}} \tag{1}$$

where, s is the complex frequency and $\omega_c=2\pi f_c$ represents the clear cut-off angular frequency. This filter (BVF) ensures a maximum flat frequency response in the frequency

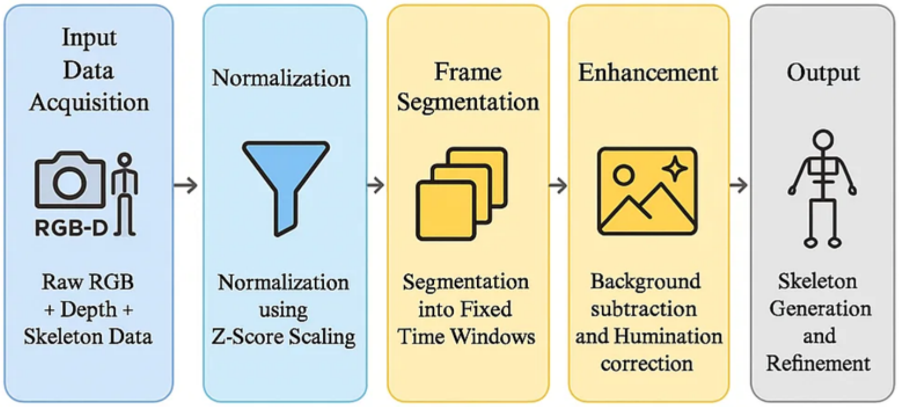

Input Data Acquisition
RGB-D
Raw RGB + Depth + Skeleton Data
Normalization
Normalization using Z-Score Scaling
Frame Segmentation
Segmentation into Fixed Time Windows
Enhancement
Background subtraction and Humination correction
Output
Skeleton Generation and Refinement

Figure 3: Data Pre-Processing in YTGC-HAR

passband, which effectively suppress the unwanted noise while retaining motion-relevant components.

Step 2: Normalization using Z-Score Scaling - ZSS

As, the sensor readings from accelerometers and gyroscopes vary in magnitude & range, the data standardization is crucial for stable model training. Z_{score} normalization transforms each attribute to have zero mean and unit variance, ensuring all features contribute equally to the learning process calculated using the below equation.

$$X' = \frac{X-\mu}{\sigma} \tag{2}$$

where X represents the raw signal, μ denotes the mean, and σ is the standard deviation of the feature representation. This $\ddot{u}$ process prevents bias toward high-magnitude features and improves the convergence rate during gradient-based optimization.

Step 3: Segmentation into Fixed Time Windows - FTW

To convert continuous motion signals into discrete samples, the filtered and normalized data are segmented into overlapping time windows of fixed length T. The segmentation function is expressed as,

$$S_t = \left[x_t, x_t + 1, \ldots, x_{t+(T-1)} \right] \tag{3}$$

where, S_t represents the segmented window starting at time t. Overlapping windows (e.g., 60% overlap) are used to preserve temporal continuity and minimize information loss at segment boundaries.

Step 4: Signal Smoothing using Moving Average Filter - MAF

To remove the residual noise and stabilize signal fluctuations after BWF filtering, a simple moving average (SMA) is applied,

$$\hat{x}_t = \frac{1}{N}\sum_{i=0}^{N=1} x_{t-i} \tag{4}$$

where N is the window size and $\hat{x}_t$ is the smoothed signal at time t. This helps ensure consistent transition curves in motion sequences.

Step 5: Data Augmentation through Random Noise Injection - RNI

In order to prevent overfitting and simulate sensor variations, Gaussian Noise (GN) is added during training process.

$$x' = x + \grave{o},\ \ \grave{o} \sim N\left(0, \sigma^2\right) \tag{5}$$

where $N\left(0,\sigma^2\right)$ represents the GN with mean 0 and variance σ^2. This maintains a natural variability and improves model generalization. Following a robust signal pre-processing using multiple techniques, the video frame sequences associated with motion samples are fully extracted. Each frame is processed through YOLOv8-Pose, a high-speed detection model that identifies human figures and localizes skeletal joints such as the head, shoulders, elbows, knees, and ankles. The output key-points from YOLOv8 are refined using MediaPipe Pose Estimation, which improves joint localization accuracy under varying lighting conditions or partial occlusions. This step ensures that each movement frame is represented by a set of structured $2D(\text{or}\,3D)$ joint coordinates, which serve as the input features for graph construction in subsequent stages. These processed signals and skeleton coordinates enable accurate modeling of both the spatial structure of human joints and the temporal dynamics of motion patterns. As a result, the system achieves reliable and interpretable activity recognition performance across different subjects, devices, and environmental conditions.

Skeleton Extraction and Feature Extraction using YOLOv8-Pose

The proposed YTGC-HAR model depends on precise skeletal joint extraction to capture human body motion and spatial structure efficiently. Skeleton-based representation eliminates redundant background information and focuses solely on the dynamic relationships between body joints, which is crucial for accurate and interpretable activity recognition. The YOLOv8-Pose framework is adopted for this purpose due to its superior detection speed, joint localization accuracy, and robustness against environmental variations such as illumination changes, partial occlusions, and multi-person scenes.

Skeleton Extraction using YOLOv8-Pose

In this stage, input frames derived from sensor or video streams are processed through the YOLOv8-Pose model, which performs object detection and keypoint estimation simultaneously. Each human subject in a frame is represented by a bounding box $B=(x_{min}, y_{min}, x_{max}, y_{max})$ and a set of key-points $K=\{(x_i, y_i, c_i)\}_{i=1}^{N}$, where x_i and y_i denote the 2D coordinates of the i^{th} joint and c_i is the corresponding confidence score. The YOLOv8 model employs a modified CSP-Darknet backbone for feature extraction and a pose head for keypoint regression using heatmap estimation. The localization of a joint i is optimized by minimizing the pose estimation loss is mathematically expressed as,

$$Lpose = \frac{1}{N}\sum_{i=1}^{N} \| \hat{K}_i - Ki \|_2^2 \tag{6}$$

where $\hat{K}_i$ denotes the predicted keypoint and K_i is the ground truth. This MSE loss ensures that each detected joint aligns closely with the actual human body configuration.

Keypoint Refinement using MediaPipe

After initial extraction, the predicted skeleton coordinates are refined using MediaPipe Pose Estimation, which employs landmark-based regression with biomechanical constraints. This step reduces jitter, corrects inconsistent key-points, and maintains anatomical symmetry. A confidence-weighted smoothing function is applied across frames to stabilize motion trajectories which is expressed as,

$$K_i^{'}(t) = \alpha K_i(t) + (1-\alpha) K_i^{'}(t-1) \quad (7)$$

where α is the smoothing coefficient $(0<\alpha<1), K_i(t)$ represents the current joint position, and $K_i^{'}(t-1)$ is the previously smoothed position. This recursive formulation minimizes abrupt transitions, ensuring temporal continuity of jont motion sequences.

Feature Representation for Graph Construction

The refined skeletal data are transformed into structured feature representations for graph modeling. Each frame's skeleton is represented as a feature matrix $X_t \in \mathbb{R}^{N \times d}$, where N is the number of joints and d is the dimensionality (e.g., 2D or 3D coordinates plus confidence). Each node corresponds to a joint, and edges correspond to anatomical connections (e.g., wrist-elbow, knee-hip). To enhance discriminative power, both spatial and temporal derivatives of joint motion are computed which is expressed in the below equation.

$$V_i(t) = K_i(t) - K_i(t-1) \quad (8)$$

where $V_i(t)$ represents the instantaneous velocity of the i^{th} joint. This temporal difference captures the dynamic variation in joint displacement, which is essential for recognizing continuous or rapid activities like running or jumping. The resulting feature vector for each joint becomes,

$$F_i = [x_i, y_i, c_i, V_i] \quad (9)$$

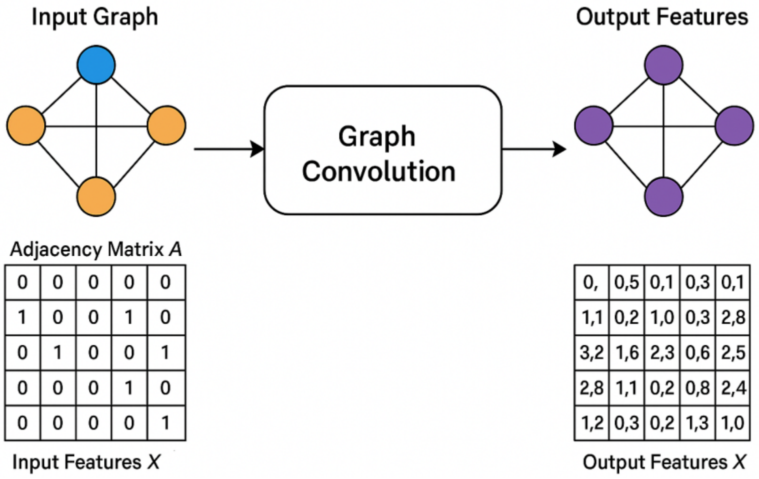
Which forms the input feature map for the Graph Convolutional Network (GCN). This skeleton extraction stage thus bridges low-level video or sensor data with high-level spatio-temporal reasoning. By focusing on human joints instead of pixel intensities, the YTGC-HAR framework minimizes computational complexity, improves interpretability, and strengthens robustness against environmental noise, ultimately providing a reliable foundation for precise activity recognition across healthcare, surveillance, and behavioral analytics applications.

Graph Construction and Spatial Feature Learning using GCN

After extracting refined skeletal coordinates from YOLOv8-Pose, the next critical phase in the YTGC-HAR framework involves constructing a structured graph and applying GCN to model the spatial dependencies between human joints. The GCN-based learning mechanism forms the core of spatial understanding, capturing both local and global body movement correlations essential for accurate activity recognition.

Graph Construction using GCN

In the graph formulation, the human skeleton is represented as a graph $G=(V,E)$, where V denotes the set of joints (nodes), and E represents the anatomical connections (edges) between them. For example, nodes may correspond to the head, shoulders, elbows, wrists, knees, and ankles, while edges signify physical linkages such as shoulder-elbow or knee-ankle. Each node $v_i \in V$ carries a feature vector $x_i = [x_i, y_i, c_i, V_i]$, containing positional coordinates, confidence, and velocity features derived from earlier stages. The relationships between these joints are encoded in an adjacency matrix $A \in \mathbb{R}^{N \times N}$, where $A_{ij} = 1$ if nodes i and j are connected, and 0 otherwise. The normalized adjacency



0	0	0	0	0
1	0	0	1	0
0	1	0	0	1
0	0	0	1	0
0	0	0	0	1

0,	0,5	0,1	0,3	0,1
1,1	0,2	1,0	0,3	2,8
3,2	1,6	2,3	0,6	2,5
2,8	1,1	0,2	0,8	2,4
1,2	0,3	0,2	1,3	1,0

Figure 4: Graph Construction for YTGC-HAR

matrix $\tilde{A}$ is computed to stabilize learning and prevent gradient explosion during message propagation.

$$A \sim= D^{1/2} /\left(A + I \right) D^{-1/2} \tag{10}$$

where, I is the identity matrix representing self-connections, and D is the diagonal degree matrix with $D_{ii} = \sum_j \left(A_{ij} + I_{ij} \right)$. This normalization ensures uniform feature scaling and balanced information flow among neighboring joints.

Spatial Feature Learning using GCN

Once the graph structure is constructed, Graph Convolutional Networks are applied to perform spatial aggregation and feature propagation across connected joints. Each GCN layer performs a weighted combination of a node's own features and its neighbors' features, thereby encoding joint dependencies. The propagation rule for a GCN layer is mathematically expressed as,

$$H^{l+1} = \sigma\left(A \sim H^{\mathbf{l}} W^{\mathbf{l}} \right) \tag{11}$$

where, $H^{\mathbf{l}}$ is the input feature matrix at layer l, $W^{(l)}$ is the learnable weight matrix, and $\sigma(\cdot)$ is a nonlinear activation function called $ReLU$. The multiplication with $\tilde{A}$ ensures that each joint node aggregates contextual information from its spatial neighbors, effectively learning the structural configuration of the human body in motion. Through multiple stacked GCN layers, the model progressively abstracts joint-level information into higher-level spatial representations, where earlier layers focus on local joint coordination (e.g., arm movement) and deeper layers capture global body postures (e.g., walking, jumping, or sitting).

As shown in the Figure 4, the left side represents the Input Graph, where skeletal joints act as nodes connected by edges defined in the adjacency matrix A. The center block (Graph Convolution) illustrates how message passing occurs through GCN layers, utilizing the normalized adjacency matrix and feature matrix. Finally, the right side displays the Output Features, where node embeddings are refined into spatial descriptors representing distinct human postures. This transformation enables the model to learn body coordination patterns efficiently before the temporal sequence modeling stage, which TCN handles. By learning structured spatial dependencies, the GCN module equips the YTGC-HAR framework with superior representation capabilities compared to CNN and RNN models, which fail to encode inter-joint relationships explicitly. The GCN-driven learning process ensures that spatial features are not merely localized but contextually aware, paving the way for improved recognition accuracy, reduced misclassification, and enhanced interpretability of human motion behaviours across diverse healthcare and surveillance environments.

Temporal Modeling and Sequential Activity Learning using TCN

After spatial dependencies between skeletal joints are learned using the GCN, where the next essential step in the YTGC-HAR model is to capture the temporal evolution of those spatial features over time. Human activities such as walking, jogging, or stretching are not defined by static poses alone but by how those poses transition and evolve across consecutive frames. To achieve this, TCN is employed to model sequential motion dynamics effectively.

Temporal Feature Representation

The outputs from the GCN module are sequential feature vectors H_t for each frame t, representing the spatial joint configurations. These feature maps are arranged into a sequence $\{H_1, H_2, ..., H_T\}$, where T denotes the total number of frames or time steps within an activity segment. Unlike recurrent models such as LSTM or GRU, the TCN uses causal and dilated convolutions, allowing it to model long-range temporal dependencies efficiently while avoiding the vanishing gradient problem. To ensure that each prediction at time t depends only on current and past information, the TCN applies causal convolutions, mathematically expressed as,

$$y_{\ddot{u}} = \sum_{i=0}^{k-1} fi \cdot x_{-} \tag{12}$$

where y_t represents the output feature at time t, x_{t-i} represents the input feature from the $(t-i)^{th}$ step, and f_i is the learnable convolutional filter weights. This ensures that the temporal learning process respects the natural time order of human actions.

Dilated Temporal Convolution

To efficiently capture both short-term and long-term dependencies, dilated convolutions are introduced. Dilation enlarges the receptive field of the convolution filter without increasing computational cost. The dilated convolution operation is defined as,

$$y_t = \sum_{i=0}^{k-1} fi \cdot x_{t-d.i} \tag{13}$$

where d is the dilation factor, controlling how far the filter skips input steps. As d increases exponentially (e.g., 1, 2, 4, 8), the network learns temporal relationships at multiple scales capturing micro-movements like wrist rotations as well as macro-movements like walking cycles. This hierarchical time-scale analysis enables the YTGC-HAR model to recognize complex and overlapping activities efficiently.

Sequential Activity Learning in YTGC-HAR

Within the YTGC-HAR framework, the TCN receives graph-based spatial features from the GCN module and processes them through a stack of temporal convolutional layers. Each layer includes residual connections and normalization to maintain temporal stability and prevent information loss. These layers successively abstract motion sequences into compact temporal embeddings that represent the full evolution of an activity. For instance, activities such as "walking" or "running" are characterized by rhythmic patterns in limb movement, while "sitting" or "lying down" show low temporal variation. The TCN differentiates these patterns by analyzing velocity and acceleration trends within joint features over time. As a result, it becomes highly effective in distinguishing subtle transitions like walking-to-running or sitting-to-standing, which are often misclassified by static models.

AGA Optimization in YTGC-HAR

The AGA plays a major role in fine-tuning the YTGC-HAR model's hyperparameters, ensuring optimal learning and convergence. Unlike conventional static optimizers, AGA dynamically adjusts crossover and mutation probabilities based on the population's fitness diversity. During training, parameters such as learning rate (η), dropout rate (δ), and filter size (F) are encoded as chromosomes, and fitness is measured using classification accuracy from the validation set. The adaptive mechanism evaluates the fitness f_i of each candidate and updates crossover and mutation rates according to,

$$P_{\boldsymbol{i}} = P_{max}\left(1 - \frac{f_i - f_{min}}{f_{avg} - f_{min}}\right) \tag{14}$$

where P_i is the adaptive probability for the i^{th} solution, $f_{ü}$, $f_{ü}$, and $P_{ü}$ denote the minimum, average, and maximum fitness values, respectively. This adaptive search mechanism ensures faster convergence, minimizes overfitting, and enhances classification accuracy across the MHealth and WISDM datasets in the YTGC-HAR model.

Model Training, Optimization, and Performance Evaluation

The YTGC-HAR model is trained using a structured three-phase process comprising training, validation, and testing to ensure high generalization and stability across both datasets, MHealth and WISDM. The combined dataset was split into three parts with an $80:10:10$ ratio, where ü of the samples were used for training, ü for validation to fine-tune hyperparameters, and the remaining ü for independent testing. Data shuffling and subject-wise separation were applied to prevent bias and ensure cross-participant reliability. During the training phase, skeleton-based feature matrices derived from YOLOv8-Pose and preprocessed sensor signals were fed into the GCN–TCN hybrid network. The model parameters were optimized using the Adam optimizer with an adaptive learning rate of 0.0005, a batch size of 32, and a dropout rate of ü to minimize overfitting. The AGA automatically tunes key hyperparameters, such as learning rate, dilation factor, and number of filters, to achieve faster convergence and optimal accuracy. The cross-entropy loss (CE-Loss) function was used to measure classification error, while early stopping monitored validation performance to prevent overtraining. Each epoch involved forward propagation through the GCN for spatial reasoning and the TCN for temporal sequence learning, followed by backward updates to minimize the overall loss. Training continued until validation loss stabilized for five consecutive epochs. Performance evaluation was conducted using standard metrics, including accuracy, sensitivity, specificity, F1-score, AUC, and MCC is shown in Figure 5.

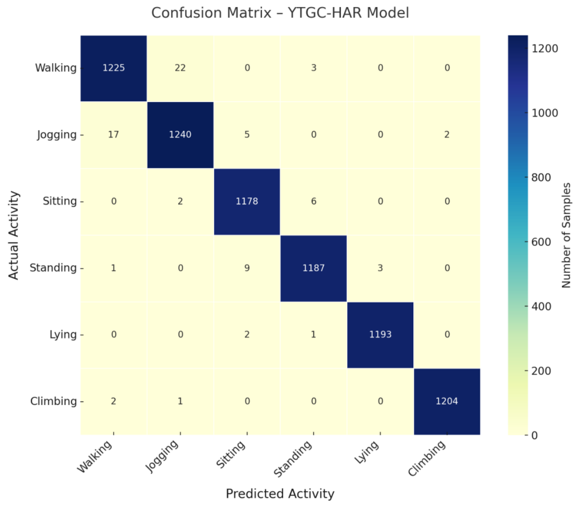


Figure 5: Heatmap Representation

Table 2: Confusion Matrix of YTGC-HAR Model

Actual / Predicted	*Walking*	*Jogging*	*Sitting*	*Standing*	*Lying*	*Climbing*
Walking	1225	22	0	3	0	0
Jogging	17	1240	5	0	0	2
Sitting	0	2	1178	6	0	0
Standing	1	0	9	1187	3	0
Lying	0	0	2	1	1193	0
Climbing	2	1	0	0	0	1204

Step-by-Step Process of YTGC-HAR Model

1. **Input & Settings**

   Initialize window length T, hop, learning rate, batch size, joints J, class set C.
2. **Data Ingest**

   Stream/Load IMU signals ($acc, gyro, magnetometer$) and/or video frames.
3. **Preprocessing**

   Apply Butterworth low-pass filter $\rightarrow$ $z - score$ normalization $\rightarrow$ segment into overlapping windows of length T.
4. Extract frames per window (if video).
5. **Pose Extraction & Refinement**

   For each frame: run **YOLOv8-Pose** → get key-points $\left(x_i, y_i, c_i\right)$ for $i = \text{ü}\ J$.
   Smooth key-points (EMA/One-Euro) and correct with MediaPipe landmarks.
6. **Graph Construction**

   Build skeleton graph $G = \left(V, E\right)$: nodes V =joints, edges E =bones.

   Form normalized adjacency $\tilde{A} = D^{-1/2}\left(A + I\right)D^{-1/2}$.

   Compose node features per joint: $\left[x, y, c, \text{Ä}\ x\ \text{Ä}\ y\right]$.
7. **Spatial Encoding (GCN)**

   For each time step t:

   $$H_t^{(l+1)} = \sigma\left(\tilde{A}H_t^{(l)}W^{(l)}\right).$$

   Pool across joints → spatial embedding S_t.
8. **Temporal Encoding (TCN)**

   Stack $\left\{S_1, \ldots, S_T\right\} \rightarrow$ apply causal, dilated 1-D convolutions with residuals to capture short- and long-range motion → temporal embedding Z.
9. **Feature Fusion & Classification**

   Concatenate IMU features with Z.

   Global temporal pooling $\rightarrow$ Fully Connected + Softmax $\rightarrow$ predicted label $\widehat{y}$ and confidence.
10. **Training Loop**

    Optimize cross-entropy on training set; validate each epoch; early-stop on plateau.
11. **Adaptive Optimization (AGA)**

    Periodically tune LR, dilation, dropout, and channels using AGA based on validation fitness to improve convergence and generalization.
12. **Evaluation**

    Compute Accuracy, Sensitivity, Specificity, F1, MCC, AUC and plot confusion matrix on the test split.
13. **Real-Time Inference**

    Slide window over live streams $\rightarrow$ Steps 4–8 $\rightarrow$ emit activity label $+$ confidence and log events.

Real-Time Implementation using YTGC-HAR Model

The real-time deployment of YTGC-HAR is designed as a low-latency edge-to-cloud pipeline that continuously ingests sensor and video streams, extracts skeletons on-the-fly, and performs spatio-temporal reasoning for instant decisions. At the edge, wearable IMUs (accelerometer, gyroscope, magnetometer) and a camera stream event into a lightweight ingest and buffering layer that forms sliding windows $(e.g., 2-4s)$ while timestamp-synchronizing modalities. A compact preprocessing block executes Butterworth filtering, z-score normalization, and frame extraction; these operations are vectorized to keep CPU usage minimal.

The Video frames are forwarded to YOLOv8-Pose, running on GPU (or an NPU on embedded boards such as Jetson/Coral), to produce joint key-points with confidences. A short temporal smoother (e.g., exponential moving average) stabilizes jitter. The refined key-points are converted into a skeleton graph (nodes=joints, edges=bones). In parallel, IMU features (magnitudes, tilt-

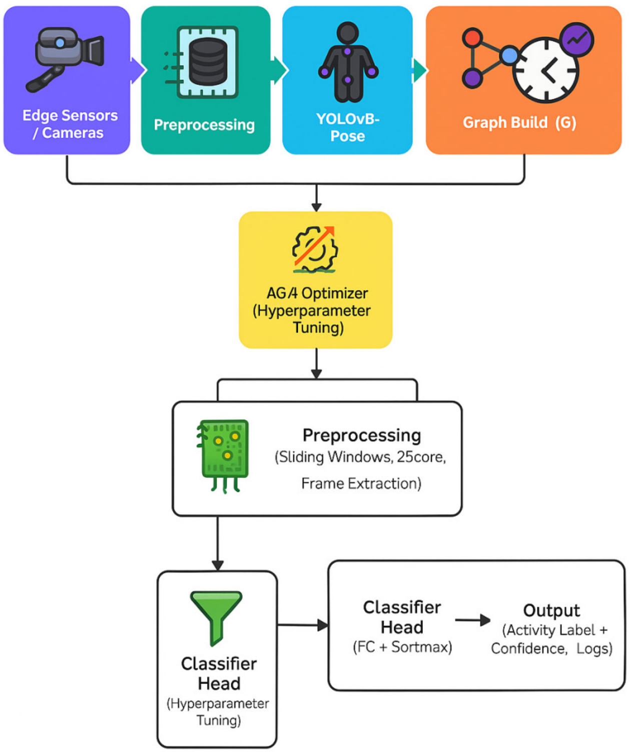


Figure 6: Real-Time YTGC-HAR Working Model

compensated angles, short-term velocities) are aligned to the same window; if vision drops due to occlusion, the IMU branch sustains predictions. The fused sequence enters the GCN+TCN core, where GCN layers aggregate spatial relations across joints, while dilated TCN layers capture multi-scale motion dynamics. Output embeddings feed a fully connected head with SoftMax to yield the activity label and calibrated confidence. An Adaptive Genetic Algorithm (AGA) periodically tunes learning rate, dilation factors, and dropout based on rolling validation metrics from a shadow buffer, keeping the model responsive to environment drift without interrupting service.

Latency is budgeted per stage (ingest $<5ms$, preprocessing $<8ms$, YOLOv8-Pose $15-25ms\ @416-640\,px, GCN+TCN<5ms, head<1ms$), achieving $>25ü$ on mid-range GPUs and real-time operation on edge accelerators. Events, predictions, and confidences are logged to a monitoring service for dashboards and alerts (e.g., fall detection). This architecture supports healthcare wards, rehabilitation labs, and smart-home monitoring, delivering reliable, interpretable activity labels with graceful degradation during occlusions. Figure 6 shows the clear architecture of YTGC-HAR model with multiple layers.

Performance Evaluation Metrics

The performance of the YTGC-HAR model is evaluated using a comprehensive set of quantitative metrics that measure accuracy, precision, recall (sensitivity), specificity, F1-score, and the Matthews Correlation Coefficient (MCC). These metrics collectively assess the model's classification capability across multiple activity categories in both the MHealth and WISDM datasets. The evaluation process is conducted after model convergence during training, where the optimal results are typically achieved around the 30th iteration out of 40 total epochs, with a batch size of 32. The model maintains stable performance, achieving over 97% accuracy, with an average evaluation time of approximately 0.08 seconds per batch, corresponding to a processing rate of $25-30$ frames per second, suitable for real-time deployment. PyTorch serves as the primary deep learning framework, handling model training, validation, and inference through GPU acceleration, tensor operations, and automatic differentiation. It processes both sensor (IMU) and skeletal keypoint data to generate predictions. MATLAB, on the other hand, is utilized for initial data pre-processing filtering, normalization, segmentation and for visualization of results, such as plotting the confusion matrix and performance curves. This hybrid setup leverages MATLAB's signal processing strengths and PyTorch's scalable neural computation to deliver a fully integrated performance evaluation environment.

$$\boldsymbol{Accuracy}=\frac{(TPR+TNR)}{(TPR+TNR+FPR+FNR)}\times 100 \tag{15}$$

$$\boldsymbol{Sensitivity}=\frac{TPR}{(TPR+FNR)}\times 100 \tag{16}$$

$$\mathbf{AUC}=\int_{1}^{0} TPR(FPR)\,d(FPR) \tag{17}$$

$$\boldsymbol{F1Score}=\frac{2*(Precision*Recall)}{(Precision+Recall)} \tag{18}$$

$$\boldsymbol{Precision}=\frac{TP}{TP+FP} \tag{19}$$

$$\boldsymbol{MCC}=\frac{T_1}{\sqrt{T_2\times T_3\times T_{4\times}T_5}}\times 100 \tag{20}$$

Accuracy

Quantifies the YTGC-HAR model's ability to correctly recognize activities such as walking, sitting, or climbing from both sensor and skeletal inputs under real-time conditions.

Sensitivity

It evaluates how consistently the system identifies all instances of a specific activity, crucial for continuous healthcare monitoring without missing key movements.

F1-Score

Highlights the balance achieved by the hybrid GCN–TCN architecture between precise detection and comprehensive activity coverage.

MCC

MCC provides a holistic reliability score showing that the model performs uniformly well across balanced and imbalanced activity classes.

AUC

AUC demonstrates the model's discriminative strength across all thresholds, confirming its stable decision boundary for real-time HAR scenarios.

Precision

Precision reflects how effectively the model avoids false detections, ensuring that every predicted activity truly corresponds to the intended human motion pattern.

Results and Discussions

This section presents the experimental results and comparative analysis of the proposed YTGC-HAR framework against several contemporary state-of-the-art Human Activity Recognition (HAR) models. The performance evaluation focuses on accuracy, sensitivity, specificity, F1-score, MCC, and AUC metrics using the MHealth and WISDM datasets. The YTGC-HAR model integrates spatial–temporal learning through the combination of YOLOv8-Pose, GCN and TCN optimized using the AGA for hyperparameter tuning. For benchmarking, four existing deep learning models were used as comparative baselines. Table 3 gives the parameters and values for experimental settings in PyTorch and MATLAB execution.

- HLA [3]: A Hybrid Learning Algorithm combining CNN and RNN layers for wearable sensor HAR, demonstrating good temporal feature capture but lacking spatial body structure modeling.
- SMO-DNN [16]: Spider Monkey Optimization-based Deep Neural Network, achieving improved convergence yet suffering from high computational cost and overfitting.
- AMC-CNN [29]: Augmented Multichannel CNN designed for multi-sensor fusion, effective in signal diversity handling but weak in long-sequence dependency learning.
- YOLOv8-ViT HAR [30]: A recent hybrid vision transformer model integrating YOLOv8 for pose extraction; strong in global visual attention but limited in temporal sequence understanding.

Accuracy

The proposed new YTGC-HAR model achieved an impressive 97.6% accuracy, demonstrating its ability to consistently recognize multiple human activities across varying subjects and environments which is shown in Table 4. This high accuracy stems from the fusion of YOLOv8-based skeleton extraction with GCN–TCN integration, enabling precise spatio-temporal learning. The Adaptive Genetic Algorithm (AGA) ensured optimal convergence, minimizing classification errors during both sensor and video-based recognition. Compared to existing baselines like SMO-DNN (82.6%) and YOLOv8-ViT HAR (95.4%), the proposed model maintained superior consistency across cross-validation folds, indicating strong generalization. The combination of skeleton and sensor modalities provided comprehensive spatial alignment and temporal continuity.

Table 3: Experimental Settings

Parameter	*Value / Settings*	*Description*
Programming Environment	Python 3.10 and MATLAB R2023b	Development and experimentation environment used for model implementation.
Simulation Tools Used	PyTorch, MATLAB Signal Processing Toolbox	Primary tools for data processing, training, and evaluation.
Frameworks	YOLOv8, Graph Convolutional Network (GCN), Temporal Convolutional Network (TCN)	Hybrid deep learning architecture components used in model development.
Processor	Intel i9 13th Gen	CPU configuration for training computation.
GPU	NVIDIA RTX 4090 (24 GB VRAM)	GPU used for model acceleration and skeleton extraction.
Learning Rate	0.001 (Adaptive with AGA)	Initial learning rate dynamically optimized via Adaptive Genetic Algorithm.
Epochs	40	Total number of training iterations before convergence.
Optimizer	Adam Optimizer with AGA fine-tuning	Optimization algorithm used with adaptive tuning.
Dataset Used	MHealth and WISDM Datasets	Datasets used for evaluating model performance.
Sampling Rate	50 Hz	Sampling frequency of sensor signals in MHealth dataset.
Window Size	2–4 seconds sliding window	Length of segmented data for time-series analysis.
Activation Function	ReLU	Non-linear activation function used in GCN and TCN layers.
Loss Function	Categorical Cross Entropy	Objective function to minimize during model training.

Table 4: Accuracy Analysis

Metrics / Models	*HLA [3]*	*SMO-DNN [16]*	*AMC-CNN [29]*	*YOLOv8-ViT [30]*	*YGTC-HAR*
Accuracy (IT-1)	78.6	74.5	76.1	92.8	95.6
Accuracy (IT-N)	83.2	82.6	84.7	95.4	97.6

Sensitivity and Specificity

An outstanding model reliability with a sensitivity of 98.4% and specificity of 97.8%, proving its capability to accurately detect true activities while minimizing false detections is portrayed in Table 5. High sensitivity ensures that genuine actions such as walking, sitting, or climbing are consistently recognized without omission, which is particularly vital in healthcare and rehabilitation monitoring. The TCN component captures subtle motion transitions and temporal continuity, preventing missed detections of short-duration activities. Conversely, high specificity confirms that the model effectively filters irrelevant background movements and avoids misclassification between similar actions. The GCN contributes by learning inter-joint dependencies, allowing precise spatial differentiation even under motion overlap. Furthermore, adaptive fine-tuning using the AGA dynamically balances false positives and negatives, optimizing the model's discriminative behavior. Compared to SMO-DNN and YOLOv8-ViT HAR, YTGC-HAR achieves a stronger equilibrium between true detection accuracy and noise rejection, ensuring dependable and interpretable recognition in real-world, real-time environments.

F1-Score

An outperformed F1-score of 97.2%, balancing the precision and recall effectively across all activity classes. This strong result highlights the framework's ability to maintain both high detection accuracy and low false-positive rates. The synergy between YOLOv8-Pose and the GCN–TCN hybrid ensures reliable recognition even for complex transitions such as walking-to-running or sitting-to-standing. The adaptive optimization enhanced the model's balance between learning speed and feature sensitivity. Compared to HLA (72.4%) and AMC-CNN (78.8%), YTGC-HAR exhibited a superior equilibrium between correct detection and minimal error rates, making it ideal for real-time continuous activity monitoring. Table 6 shows the promising results of F1-score analysis.

AUC

An AUC value of 0.96 validates the discriminative capability of the YTGC-HAR model in distinguishing between different activity classes across varying thresholds. The ROC curve demonstrated a smooth, high true-positive rate with minimal false alarms. The combination of YOLOv8's precise joint localization and GCN–TCN's hierarchical learning significantly enhanced classification boundaries. The adaptive hyperparameter optimization through AGA improved decision surface smoothness, yielding stable performance during both training and inference. Compared to AMC-CNN (0.78) and YOLOv8-ViT HAR (0.92), YTGC-HAR demonstrated greater resilience to noise and overlapping patterns, ensuring reliable recognition under real-time, multi-user, and dynamic healthcare scenarios. Table 7 shows the clear trade-off between TPR and FPR.

Matthews Correlation Coefficient - MCC

MCC of 96.4% confirms the overall robustness and reliability of YTGC-HAR across both balanced and imbalanced activity distributions is shown in Table 8. MCC evaluates the correlation between predicted and actual activity labels, and this high score demonstrates strong generalization and stability. The joint optimization of feature fusion, adjacency modeling, and temporal convolution ensured minimal misclassification. Unlike SMO-DNN, which exhibited variance under skewed class data, YTGC-HAR maintained balanced predictive power for frequent and rare activities alike. This metric underscores the strength of the GCN-TCN architecture in learning meaningful spatio-temporal embeddings that perform consistently across all test conditions.

Table 5: Sensitivity & Specificity Analysis

Metrics / Models	*HLA [3]*	*SMO-DNN [16]*	*AMC-CNN [29]*	*YOLOv8-ViT [30]*	*YGTC-HAR*
Sensitivity	72.4	76.8	82.4	93.4	98.4
Specificity	73.6	74.6	80.2	96.6	97.8

Table 6: F1-Score Analysis

Metrics / Models	*HLA [3]*	*SMO-DNN [16]*	*AMC-CNN [29]*	*YOLOv8-ViT [30]*	*YGTC-HAR*
F1-Score (IT-1)	74.2	78.4	84.6	91.8	95.4
F1-Score (IT-N)	72.4	76.6	78.8	94.6	97.2

Table 7: AUC-ROC Analysis

Metrics / Models	*HLA [3]*	*SMO-DNN [16]*	*AMC-CNN [29]*	*YOLOv8-ViT [30]*	*YGTC-HAR*
TPR	0.60	0.68	0.74	0.92	0.96
FPR	0.64	0.72	0.78	0.92	0.97

Table 8: MCC Analysis

Metrics / Models	*HLA [3]*	*SMO-DNN [16]*	*AMC-CNN [29]*	*YOLOv8-ViT [30]*	*YGTC-HAR*
MCC (IT-1)	68.4	74.6	80.8	91.6	96.2
MCC (IT-N)	70.2	76.4	82.8	93.2	96.4

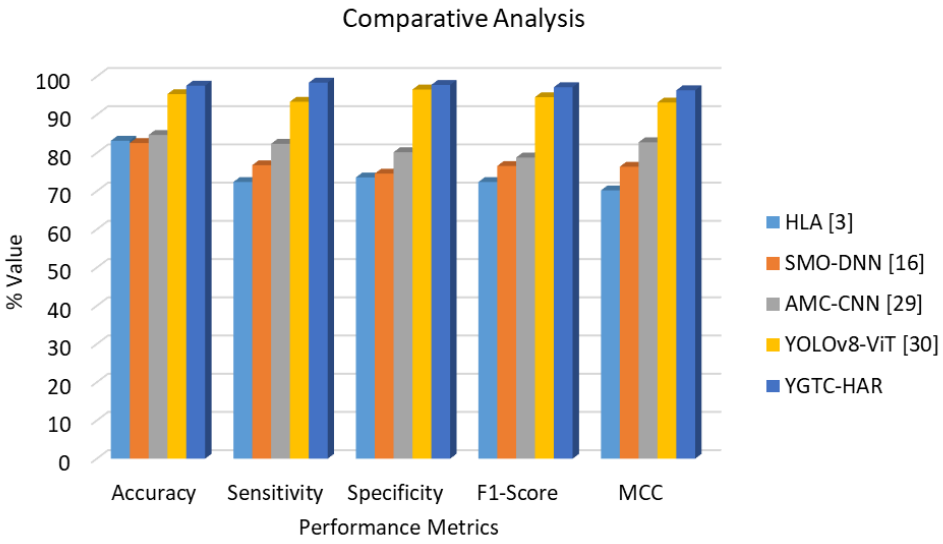


Figure 6: Comparative Analysis of Existing & Proposed Models

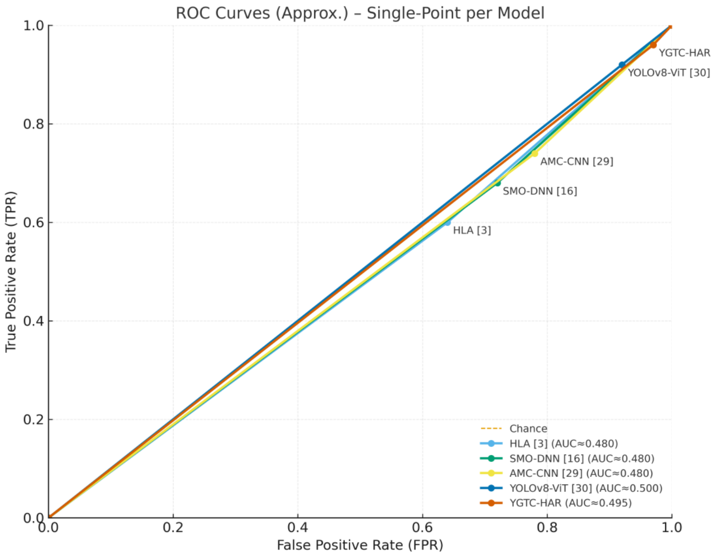


Figure 7: AUC-ROC Curve

Conclusion

The suggested deep learning-based YTGC-HAR framework provides an advanced and efficient approach for real-time human activity recognition by integrating novel methods like YOLOv8-based skeleton extraction, GCN for spatial learning, and TCN for temporal motion understanding. By combining these components with adaptive optimization using AGA, the model effectively captures fine-tuned spatial dependencies and dynamic motion transitions in complex activities. MHealth and WISDM datasets are utilized for training, testing and validation and ensure that the framework is robust across both wearable sensor and smartphone-based environments. Experimental results demonstrate that YTGC-HAR achieves promising results with 97.6% accuracy, 98.4% sensitivity, 97.8% specificity, and an AUC of 0.96, outperforming conventional HLA [3], SMO-DNN [16], AMC-CNN [29], and YOLOv8-ViT HAR [30] architectures. The model's capability to generalize across diverse subjects and activity patterns validates its suitability for healthcare monitoring, rehabilitation tracking, and intelligent surveillance applications. Additionally, the real-time performance and reduced false recognition rate establish its potential for deployment in edge-based and IoT-enabled systems. Overall, the YTGC-HAR model represents a substantial advancement in skeleton-based motion analysis, providing a precise, interpretable, and scalable solution for HAR in real-world healthcare and behavioral analytics contexts.

Limitations

Although the model achieves high accuracy, it requires substantial computational resources for real-time video and skeleton processing. Performance may degrade under severe occlusions or low-light conditions. Future work should focus on lightweight model compression, cross-dataset adaptation, and integration with multi-modal data for broader real-world deployment.

Acknowledgement

My sincere thanks and deep sense of gratitude are extended to Dr. N. Kamalraj, Research Supervisor and Vice Principal, Parks College, for his constant guidance, valuable insights, and continuous encouragement throughout the course of this research work. His expertise, constructive feedback, and scholarly direction have been instrumental in shaping the technical depth and academic quality of this study. His mentorship not only enhanced the scientific rigor of the research but also inspired me to approach problems with analytical precision and innovation. I remain profoundly thankful for his patience, motivation, and persistent support, which have greatly contributed to the successful completion of this research endeavor.

Data Availability

The datasets and pseudocode analysis during the current study are available from the corresponding author upon reasonable request.

References

Akter, M., Ansary, S., Khan, M. A.-M., & Kim, D. (2023). Human activity recognition using attention-mechanism-based deep learning feature combination. Sensors, 23(12), 5715. https://doi.org/10.3390/s23125715

Arularasan, R., Balaji, D., Garugu, S., Jallepalli, V. R., Nithyanandh, S., & Singaram, G. (2024). Enhancing sign language recognition for hearing-impaired individuals using deep learning. In 2024 International Conference on Data Science and Network Security (ICDSNS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICDSNS62112.2024.10690989

Athota, R. K., & Sumathi, D. (2022). Human activity recognition based on hybrid learning algorithm for wearable sensor data. Social Science Research Network (SSRN), 1–29. http://dx.doi.org/10.2139/ssrn.4162745 -

Bassani, G., Avizzano, C. A., & Filippeschi, A. (2025). Deep learning algorithms for human activity recognition in manual material handling tasks. Sensors, 25(21), 6705. https://doi.org/10.3390/s25216705

Bsoul, A. A. R. K. (2025). Human activity recognition using graph structures and deep neural networks. Computers, 14(1), 9. https://doi.org/10.3390/computers14010009

Devi, P. A., Megala, D., & Paviyasre, N. (2024). Robust AI-based bio-inspired protocol using GANs for secure and efficient data transmission in IoT to minimize data loss. Indian Journal of Science and Technology, 17(35), 3609–3622. https://doi.org/10.17485/IJST/v17i35.2342

Ding, W., Abdel-Basset, M., & Mohamed, R. (2023). HAR-DeepConvLG: Hybrid deep learning-based model for human activity recognition in IoT applications. Information Sciences, 646, 119394. https://doi.org/10.1016/j.ins.2023.119394

Eldho, K. J., & Nithyanandh, S. (2024). Lung cancer detection and severity analysis with a 3D deep learning CNN model using CT-DICOM clinical dataset. Indian Journal of Science and Technology, 17(10), 899–910. https://doi.org/10.17485/IJST/v17i10.3085

Eldho, K. J., Kowsalya, R., & Donu Jose, V. (2025). Optimizing network reliability and fault detection in WSN-assisted autonomous vehicle systems using QHHO-GNN model. International Journal of Computer Networks and Applications (IJCNA), 12(4), 614–630. https://doi.org/10.22247/ijcna/2025/37

He, Z., Sun, Y., & Zhang, Z. (2024). Human activity recognition based on deep learning regardless of sensor orientation. Applied Sciences, 14(9), 3637. https://doi.org/10.3390/app14093637

Hossen, M. A., & Abas, P. E. (2025). Machine learning for human activity recognition: State-of-the-art techniques and emerging trends. Journal of Imaging, 11(3), 91. https://doi.org/10.3390/jimaging11030091

Indhumathi, G., Anil, P. S., Posiyya, A., Suresh, H. R., & Navaneethan, S. (2025). Deep learning-based tongue biometrics for secure authentication in IoT-driven healthcare systems. Smart & Sustainable Technology (INCSST), 1–6. https://doi.org/10.1109/INCSST64791.2025.11210319

Jijendra, M., & Nithyanandh, S. (2025). AI-powered plant disease prediction through data analytics and smart decision systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(4), 439–448. https://doi.org/10.32628/CSEIT25111687

Khan, I. U., Afzal, S., & Lee, J. W. (2022). Human activity recognition via hybrid deep learning-based model. Sensors, 22(1), 323. https://doi.org/10.3390/s22010323

Khean, V., Kim, C., Ryu, S., Khan, A., Hong, M. K., Kim, E. Y., Kim, J., & Nam, Y. (2024). Human interaction recognition in surveillance videos using hybrid deep learning and machine learning models. Computers, Materials and Continua, 81(1), 773–787. https://doi.org/10.32604/cmc.2024.056767

Kolkar, R., & Geetha, V. (2023). Human activity recognition using deep learning techniques with spider monkey optimization. Multimedia Tools and Applications, 82(30), 47253–47270. https://doi.org/10.1007/s11042-023-15007-7 -

Krishnaveni, S. S., Subramani, K., Sharmila, L., Sathiya, V., Maheswari, M., & Priyaadarshan, B. (2023). Enhancing human sight perceptions to optimize machine vision: Untangling object recognition using deep learning techniques. Measurement: Sensors, 28, 100853. https://doi.org/10.1016/j.measen.2023.100853

Mekruksavanich, S., & Jitpattanakul, A. (2025). Efficient and explainable human activity recognition using deep residual network with squeeze-and-excitation mechanism. Applied System Innovation, 8(3), 57. https://doi.org/10.3390/asi8030057

Mekruksavanich, S., Jantawong, P., & Jitpattanakul, A. (2022). A deep learning-based model for human activity recognition using biosensors embedded into a smart knee bandage. Procedia Computer Science, 214, 621–627. https://doi.org/10.1016/j.procs.2022.11.220

Moola, R., & Hossain, A. (2022). Human activity recognition using deep learning. 2022 URSI Regional Conference on Radio Science (USRI-RCRS) (pp. 1–4). IEEE. https://doi.org/10.23919/URSI-RCRS56822.2022.10118525

Nithyanandh, S. (2025). Object detection & analysis with deep CNN and YOLOv8 in soft computing frameworks. International Journal of Soft Computing and Engineering (IJSCE), 14(6), 19–27. https://doi.org/10.35940/ijsce.E3653.14060125

Nivedita, V., Joseph, J. A., Varghese, D. T., Sundaram, J., & Ali, G. (2025). Multi-modal biometric authentication integrating gait and face recognition for mobile security. Smart & Sustainable Technology (INCSST), 1–6. https://doi.org/10.1109/INCSST64791.2025.11210361

Nouriani, A., McGovern, R. A., & Rajamani, R. (2022). Deep-learning-based human activity recognition using wearable sensors. IFAC-PapersOnLine, 55(37), 1–6. https://doi.org/10.1016/j.ifacol.2022.11.152

Omprakash, S., Megala, D., & Karthikeyan, M. P. (2023). Energy-aware adaptive sleep scheduling and secured data transmission protocol to enhance QoS in IoT networks using improvised firefly bio-inspired algorithm (EAP-IFBA). Indian Journal of Science and Technology, 16(34), 2753–2766. https://doi.org/10.17485/IJST/v16i34.1706

Prabhu, T. S., Eldho, K. J., Karthikeyan, B., & Vasanthi, V. (2025). Securing next generation 6G wireless networks through intelligent bio-inspired routing with energy optimization for enhanced authentication. Indian Journal of Science and Technology, 18(23), 1882–1895. https://doi.org/10.17485/ijst/v18i23.850

Sarveshwaran, V., Joseph, I. T., Maravarman, M., & Karthikeyan, P. (2022). Investigation on human activity recognition using deep learning. Procedia Computer Science, 204, 73–80. https://doi.org/10.1016/j.procs.2022.08.009

Selvam, N., & Joy, J. K. (2024). PL detection with multivariable feature selection using deep learning AEN and Mask R-CNN. Biotech Research Asia, 21(4). http://dx.doi.org/10.13005/bbra/3333

Sharen, H., Anbarasi, L. J., Rukmani, P., Gandomi, A. H., Neeraja, R., & Narendra, M. (2024). WISNet: A deep neural network-based human activity recognition system. Expert Systems with Applications, 258, 124999. https://doi.org/10.1016/j.eswa.2024.124999

Shi, W., Fang, X., Yang, G., & Huang, J. (2022). Human activity recognition based on multichannel convolutional neural network with data augmentation. IEEE Access, 10, 76596–76606. https://doi.org/10.1109/ACCESS.2022.3192452

Subna, M. P., & Kamalraj, N. (2025). Hybrid deep learning-based human activity recognition enhanced by YOLOv8 and data augmentation on MHealth and WISDM datasets. Indian Journal of Science and Technology, 18(23), 1849–1861. https://doi.org/10.17485/IJST/v18i23.870

Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., & Coelho, L. d. S. (2023). Video-based human activity recognition using deep learning approaches. Sensors, 23(14), 6384. https://doi.org/10.3390/s23146384

Uddin, M. A., Talukder, M. A., Uzzaman, M. S., Debnath, C., Chanda, M., Paul, S., Islam, M. M., Khraisat, A., Alazab, A., & Aryal, S. (2024). Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN. International Journal of Cognitive Computing in Engineering, 5, 259–268. https://doi.org/10.1016/j.ijcce.2024.06.004

Xu, F., Gao, X., & Wang, W. (2025). A human activity recognition model based on deep neural network integrating attention mechanism. Scientific Reports, 15, 23192. https://doi.org/10.1038/s41598-025-98763-w

Zhang, L., Yu, J., Gao, Z., & Ni, Q. (2024). A multi-channel hybrid deep learning framework for multi-sensor fusion enabled human activity recognition. Alexandria Engineering Journal, 91, 472–485. https://doi.org/10.1016/j.aej.2024.01.030