INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Hybrid Deep Learning-Based Human Activity Recognition Enhanced by YOLOv8 and Data Augmentation on MHealth and WISDM Datasets

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

1 Research Scholar (Full Time), Department of Computer Science, Park's College (Autonomous), Chinnakkarai, Tirupur-641605
2 Vice Principal, Park's College (Autonomous), Chinnakkarai, Tirupur-641605

OPEN ACCESS

Received: 09/05/2025

Accepted: 29/05/2025

Published: 18/06/2025

Citation: Subna MP, 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

* **Corresponding author**.

rehnasubi@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To propose an advanced Human Activity Recognition Framework to classify complex human activities with high accuracy and computational efficiency by integrating hybrid deep learning and optimization techniques. The model helps to build a robust system capable of maintaining high performance across diverse user behaviours and heterogeneous sensor modalities. **Methods:** Hybrid deep learning YOLOv8 extracts rapid spatial features from image and video frames. Vision Transformer (ViT) enhances the classification depth and captures long-range dependencies and global context. To ensure optimal tuning, particle swarm optimization (PSO) is applied for hyperparameter optimization, which includes batch size, learning rate, and drop ratio. Synthetic data augmentation, such as cropping, rotation, flipping, etc., is done using GAN-based generation, which helps to enhance generalization. The YOLOv8-ViT with PSO model extracts data from two benchmark datasets, WISDM and MHealth, comprising physical activities, multi-dimensional time series data, and annotated sensor readings. The proposed model is evaluated using MATLAB and Python, and the results are compared with existing models such as HLA, SMO-DNN, and AMC-CNN. **Findings:** The proposed model yields promising results with 95.4% accuracy, 96.6% sensitivity, 96.8% specificity, 0.92 AUC, 93.2% MCC, and 94.6% F1-Score, which is substantially outperforming the existing computational models for HAR with different datasets. **Novelty:** This research presents a multi-layered architecture combining YOLOv8's speed, contextual power of ViT, PSO's adaptive tuning, and GAN-based data augmentation, which offers scalable, real-time, and high precision HAR models for future healthcare and innovative sensing applications.

Keywords: Human Activity Recognition (HAR); YOLOv8; Deep Learning; Vision Transformer; Particle Swarm Optimization; Data Augmentation

1 Introduction

Human Activity Recognition (HAR) has gained attention in recent years due to the rapid growth of high-speed sensor technological advancements, especially in healthcare monitoring, fitness tracking, assisted living, and smart environments. Even with notable advancements, the prevailing models struggle with real-time performance, generalization across diverse sensor device sources, and precise recognition under complex and imbalanced datasets. Traditional machine learning and deep learning approaches rely on sequential models that simultaneously underperform in spatial and contextual dependencies. To overcome these limitations, a novel hybrid deep learning HAR framework is proposed in this study by combining the speed and precision of YOLOv8 for spatial feature extraction with ViT to capture long-range dependencies and global context, which enhances both accuracy and interpretability of the model. Additionally, PSO is integrated to tune the hyperparameter dynamically to learn the behaviours optimally. At the same time, GAN-based data augmentation improves class balance and generalization across unseen activities and user profiles based on settings. The proposed approach is validated on the MHealth and WISDM datasets, demonstrating robust performance under various sensor modalities. The model addresses the pressing need for scalable, accurate, and real-time HAR solution that suits next-generation human-centric and smart applications in multiple sectors.

Using the wireless channel state information, a WiFi-based fine-grained MS-HAR[1] method was proposed. This approach has demonstrated high accuracy without requiring any wearable sensor devices. The only shortcoming of this method is the adaptability of significant environmental sensitivity, which limits the scalability in dynamic real-world settings. A dynamic deep learning PPG-NET[2] framework to estimate heart rate using photoplethysmogram signals. Though the model is highly effective for device-independent monitoring, it is limited in HAR due to a lack of motion pattern recognition and insufficient temporal feature modeling. CRCCN-Net model[3] was developed for medical image classification using histopathological data. The tri-classifier classifies the data rigorously and achieves 91% accuracy, which stands at the top in image classification models. However, the framework mainly focused on spatial features with limited applicability in dynamic activity contexts and required temporal sequence analysis. A bio-inspired swarm-based model[4] was introduced to detect the dynamic link failure in neural networks to enhance the quality of service in sensor devices and achieve high accuracy. This model is beneficial for benchmarking sensor deployments using that protocol. Combined CNN, Conv-LSTM, and LRCN[5] for HAR were proposed to capture spatio-temporal dependencies effectively. The framework has good potential for HAR but lacks optimization and hyperparameter tuning, leading to an inefficient learning rate across iterations. A multichannel CNN[6] with data augmentation for HAR was developed to improve the dynamic transformation of the model and data diversity. The only drawback of the model is that generative methods like GAN produce rich variations in activity patterns. A hybrid DL algorithm for wearable sensor data[7] was proposed to increase the detection accuracy rate of activity monitoring using WISDM datasets. Though the accuracy is boosted, the model complexity has increased, and no optimization techniques were employed to automate the parameter selection. Deep learning-based Spider Monkey Optimization (SAM)[8] was developed to improve parameter tuning settings to achieve accuracy in HAR. Despite the model showing 80-85% accuracy with enhanced optimization, it neglects the key performance metrics like sensitivity and MCC, which are vital in HAR models. IWD-ARP[9] was proposed to detect sensor link failure rates and address the optimization issues; the approach does not address the activity classification or robustness under imbalanced data distributions. A multi-source data fusion using MLP for HAR[10] was employed to enhance the recognition accuracy and achieved 86%. Yet, the MLP lacks advanced feature extractions and fails to utilize temporal attention mechanisms, which limits the adaptability of the HAR model for complex activity sequences. Using sensor data, a hybrid CNN-RNN architecture[11] was developed for HAR. The model efficiently captures the finest spatial and sequential patterns to boost the accuracy and sensitivity. The model lacks advanced hyperparameter settings, leading to potential delays and overfitting in long activity sequence recognition. The novel multi-scale recognition algorithms[12] are reviewed to analyze the cross-subject performance in rehabilitation human activity systems across various users. The model shows limitations in incorporating the dynamic data augmentations under imbalanced activity distributions. RVSRP-OS[13] model was developed to focus more on network resilience, where the system's lifetime is boosted drastically. The model lacks direct applicability to HAR or model optimization strategies. A deep multi-task learning network[14] for HAR and user identification using smartphone sensors was designed to improve the resource utilization. Though the model achieved a potential pattern recognition method, the computational complexity increases and limits data augmentation for rare activity classes. Federated learning[15] for HAR on edge devices was discovered to promote data privacy and low-latency processing. The approach struggles with heterogeneous device capabilities and lacks advanced optimization mechanisms to stabilize the performance across all devices. The self-supervised learning approach[16] using cross-sensor masked reconstruction for HAR was proposed to reduce dependency on labeled data, where the limitations are complex activity sequence recognition with high contextual understanding. ABC[17] and EAPIFBA[18] were introduced as an optimization model to address the sensors' failure detection and energy efficiency, which are used in the current model to enhance the sensor lifetime during deployment. A dynamic CBAM[19] was proposed to improve

HAR through attention mechanism and spatial feature learning. The model lacks long-term temporal relationships, which are critical for accurate sequential activity recognition. A lightweight activity recognition[20] model was triggered to reduce the power consumption in continuous HAR. Though the model was energy efficient, the system compromises the TPR in complex multi-activity environments. Petri-Net architecture[21] was presented for real-time HAR in industrial environments, which is highly effective for structured workflows. The model lacks the flexibility to adapt to unpredictable activity patterns common in daily life. 3D-CNN[22] and DLF[23] deep learning models were proposed to detect objects and features in complex 2D and 3D images. The model lacks continuous pattern recognition and deep segmentation. AEN with Mask-R-CNN[24] was developed to detect intrinsic features in medical images with 91% accuracy. However, it does not address the temporal aspects of critical HAR. Real-life HAR model[25] was explored using smartphone sensors with ML techniques. The model works on real-time applications but lacks optimization and complex activity transitions. To overcome the limitations of the existing models, the new YOLOv8-ViT with PSO & GAN is proposed in this research work, which is examined in the next section.

1.1 Research Gap

Despite significant advancements in HAR, several research gaps persist during real-time deployments. Figure 1 shows the concept map of the research gaps identified during the literature study and state-of-the-art methods.

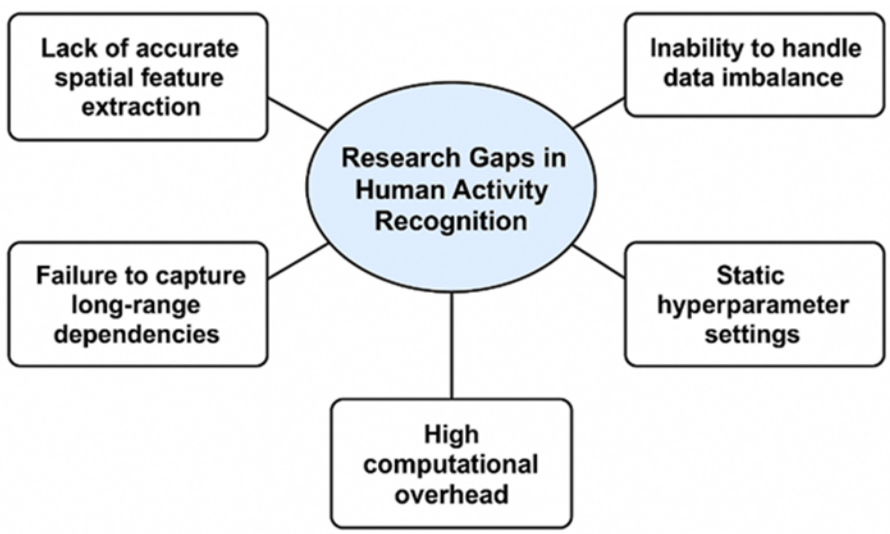

Fig 1. Concept Map of Research Gaps

To bridge the critical research gaps from existing models, the present study focused on a hybrid framework model in combination with deep learning YOLOv8-ViT with PSO & GAN to deliver a scalable, context-aware, and efficient solution to recognize human activity across various applications and sensor devices.

2 Methodology

This section presents a novel hybrid deep learning framework tailored for HAR by integrating four key computational advancements like YOLOv8, ViT, PSO, and GAN for spatial feature selection & extraction, contextual learning, hyperparameter tuning, and data augmentation. The new methodology is designed to overcome the limitations of the existing machine learning based HAR models, such as HLA[6], SMO-DNN[7], and AMC-CNN[8]. Existing models lack adequate spatio-temporal integration, rigid parameter settings, and weak generalization on imbalanced data handling. The proposed model transforms time-series sensor data from two benchmark datasets, MHealth[?] and WISDM[?], into image-like representations. The model enhances diversity & class balance, detects motion-relevant spatial regions, and processes global sequence context for activity interpretation. The batch size and dropout ratio are dynamically optimized during the model training. Unlike traditional models, this newly proposed tri-computational approach provides a novel framework for human activity recognition to enhance accuracy and real-time applicability. Each component is elaborated in this section to justify its inclusion and demonstrate how this deep learning model can effectively address the identified research gaps.

2.1 Data Acquisition and Pre-Processing

This research uses two crucial publicly available MHealth and WISDM datasets, which are ideally utilized for human activity recognition. The MHealth datasets comprise multiple readings like accelerometer, gyroscope, and magnetometer, which are collected from 10 subjects performing 12 different activities. Each sensor stream provides multivariate time-series data at a 50Hz sampling rate, totaling over 23,000 labeled activity segments. The WISDM datasets collected from smartphones and smartwatches from 36 users comprise over 1 million sensor readings for six main activities: walking, jogging, and sitting. The model generalization is ensured by splitting the datasets into a 70:30 ratio for training, testing & and validation purposes to maintain a uniform distribution of activity labels**.** Prior to model distribution the datasets are pre-processed using the below techniques

- **Noise Filtering** - To remove signal drift by using low-pass buffer worth filter
- **Segmentation** - By fixed size sliding windows with 50% overlap to preserve temporal continuity
- **Normalization** - By z-score standardization

$$\mathbf{z}score = (\mathbf{x} -)/ \tag{1}$$

where, x is raw reading, μ is mean and σ is standard deviation of the signal. The multivariate window is converted into a grayscale image representation using feature mapping techniques such as signal-image encoding by Gramian Angular Fields (GAF) and Recurrence Plots (RP). These input frames capture spatial motion signatures, enabling the new model YOLOv8 to extract location-sensitive features effectively. This multi-transformation pipeline ensures the raw time-series data, which becomes a visually interpretable & learnable input for the proposed YOLOv8-ViT with PSO & GAN hybrid model**.**

2.2 Data Augmentation using GAN

To address the class imbalance and improve the generalization in human activity recognition (HAR), the proposed model employs Generative Adversarial Networks to generate the synthetic training samples for implementation. GANs are dynamically composed of two different neural networks, G & D (Generator & Discriminator). The network is trained simultaneously in a min-max game. The generator generates samples from random noise, while the discriminator evaluates the authenticity of the samples against the real data provided. The input to the generator is a random latent vector $z \sim p_z(z)$ and optionally, real time characteristics (class labels or feature distributions). The generator transforms z into image like activity frames $G(z)$, while discriminator attempts the evaluation between real and generated synthetic samples.

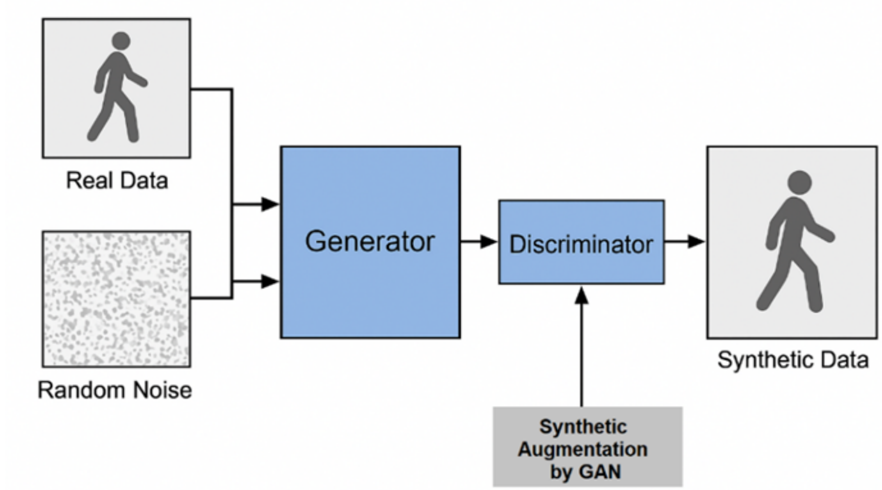

Fig 2. Data Augmentation using GAN

The overall GAN loss function is,

$$V(D,G) = E_x - p\_data(x)[\log D(x)] + E_{z-p_z(z)}[\log(1 - D(G(z)))] \tag{2}$$

where, $\min(G)\max(D)$ is to stabilize GAN and enforce better feature alignments. Here, $G(z)$ is generator synthetic data drawn from G and $D(x)$ is probability of the input (x) is real. E is the expectation operator (average over distribution). The feature matching loss function is used as,

$$L_{FM} = ||E_x[f(x)] - E_z[f(G(z))]||^2 \tag{3}$$

where, L_{FM} is feature matching loss and $f(x)$ is feature representation of real data from an intermediate layer of discriminator. Each generated frame is evaluated again using the reconstructed metric such as Structural Similarity Index (SSIM),

$$SSIM\ (x, G(z)) = \frac{(2\mu x\mu y + c1)(2\sigma_{xy} + c2)}{(\mu_x^2 + \mu_y^2 + c1)(\sigma_x^2 + \sigma_y^2 + c2)} \tag{4}$$

where, the mean of real and generated images is $\mu x\mu y$ and $c_1 and\ c_2$ are the small constants to stabilize the division when the denominator is close to zero. The quality of generated samples is measured using discriminator confidence score,

$$P_{real} = D(G(z))\,(closer\ to\ 1 = highquality) \tag{5}$$

where, P_{real} is the confidence score that the generated data is real and $D(G(z))$ is the output of the discriminator for generated sample. By augmenting the under-represented activity classes GANs ensure that the model encounters a diverse range of patterns during the training.

2.3 YOLOv8-ViT with PSO & GAN Architecture

The proposed architecture combines the strengths of YOLOv8, ViT, and PSO to build a robust, scalable, and high-performance hybrid framework for Human Activity Recognition. This is implemented through a GAN-based image representation of time-series data. A four-step process is carried out to execute this research and enhance the performance.

Step 1: YOLOv8 for spatial feature extraction
Step 2: Vision Transformer for contextual understanding
Step 3: PSO for hyperparameter optimization
Step 4: Integration of GAN augmented data

Figure 3 clearly shows the complete proposed model's architecture on how the step by step process is carried out in different forms by employing novel methods to enhance the human activity recognition by utilizing the MHealth and WISDM dataset during PyTorch and MATLAB implementation process. 30 iterations are done to test the performance of the model with different epochs.

2.3.1. YOLOv8 for Spatial Feature Extraction

You Only look Once, version 8 is employed as the first stage in the proposed model to extract the fine-grained spatial features from the input frames. The conventional backbone of YOLO processes the input frames rapidly while preserving localized motion characteristics from joints. YOLO uses anchor-free approach to predict the bounding box coordinates (x, y, w, h) directly, where x, y denotes the center coordinates of detected region and w, h represents the width and height of the feature map region. Each predicted region is scored using the object function,

$$P_{obj} = \sigma(C_{cls}) . \sigma(C_{conf}) \tag{6}$$

where, C_{cls} is the probability confidence, C_{conf} is the probability of the object presence and σ is sigmoid function. This model transforms each activity frame into dense feature map that highlights movement regions which is passed to next stage.

2.3.2. ViT for Contextual Understanding

The ViT layers receives the spatial feature maps from YOLO and start capturing the long-range dependencies and activity sequences by applying the global self-attention mechanism. The input image is divided into fixed-size non-overlapping patches and embedded into tokens. This is computed as,

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) . V \tag{7}$$

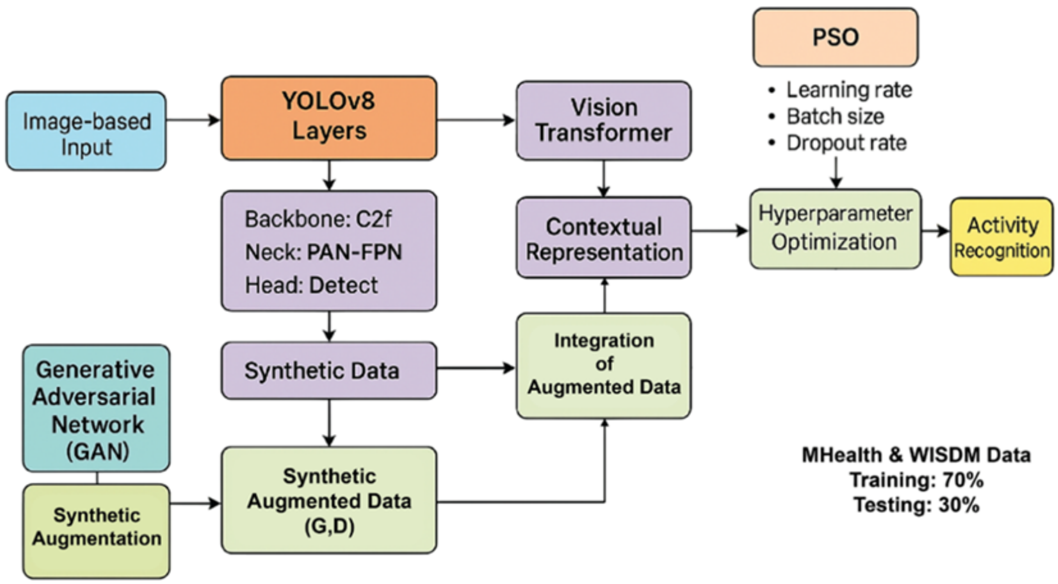

Fig 3. YOLOv8-ViT with PSO & GAN Architecture

where, Q, K, V are query, key and value metrics derive from X. d_k is the dimension of key vectors and $softmax$ is the normalization function. This mechanism allows the model to focus more on temporal regions and capturing frames in human action. The deep contextual learning is notably achieved using ViT model along with YOLO to derive the activity sequences. During real-time implementation, ViT works well by applying GSAM technique to detect long range dependencies and handle the imbalance data in a robust manner.

2.3.3. PSO for Hyperparameter Tuning and GAN for Data Augmentation

The PSO nature inspired model is integrated to optimize the parameters such as i) learning rate (α) ii) dropout rate (γ) and iii) batch size (B). Each particle in PSO represents the potential solution vector $x_i = \alpha,\ \gamma, B)$ and it is updated on every iteration with different epochs and batch sizes. The augmented data is generated by GAN is directly injected into the training stream and evaluated with the real-time data samples. The final model is trained to minimize the compose loss is represented as,

$$L_{total} = \lambda_1 L_{YOLO} + \lambda_2 L_{ViT} + \lambda_3 L_{PSOpenality} \tag{8}$$

where, L_{YOLO} is localization and classification loss from YOLOv8, L_{ViT} is cross-entropy output from ViT output, $L_{PSOpenality}$ is penalty for sub-optimal parameter configurations and $\lambda_1, \lambda_2, \lambda_3$ are balancing weights.

2.4 Training and Evaluation Strategy

The proposed YOLOv8-ViT with PSO & GAN model is trained using MHealth and WISDM data with a stratified 70:30 train-test split to ensure the equal class distributions. The training configurations are,

- Number of Iterations: 30
- Batch Size: 32 Samples per batch
- Number of Epochs: 50 per iteration
- Number of Images Trained: 14,000 Image Frames (9200 real images and 4800 GAN generated images to address class imbalance)

The predicted labels are activity class, confidence scores, feature maps, attention maps and performance metrics. The actual and predicted values are calculated by performing 2x2 confusion matrix which shows a clear classification of activity and no activity for 30 iterations with balanced sample datasets. The training and testing split with number of image frames is differ for all the iterations which improves the results of the proposed YOLOv8-ViT with PSO & GAN based hybrid deep learning model.

YOLOv8-ViT with PSO & GAN Step by Step Algorithm

1. **Input:** Time Series Data from WISDM & MHealth
2. **Begin:** Load Data
3. **Pre-Processing and Frame Conversion**

   Normalize using $z - score$ using GAF & RP
4. **Synthetic Data Augmentation using GAN**

   $D_{Train} = D_{Real}\, \upsilon\, D_{GAN}$
5. Feature extraction using YOLOv8

   $YOLOv8 = Detect\ (Movement\ Regions)$
6. **Contextual Encoding using ViT**

   $Capture\ LR\ Dependencies$
7. **Define Hyperparameter search space using PSO**

   Set learning rate, batch size and dropout ratio
8. **Train the Model**

   30 Iterations, 50 Epochs
9. **Evaluate the Metrics**
10. **Output:** Prediction and Visual Maps of HAR

Table 1. Parameter Settings for MATLAB and PyTorch

Parameters	Value / Settings	Description
Batch Size	32	Number of images per training batch
Learning Rate	0.001	Initial learning rate for optimizer
Number of Epochs	50	Epochs per training cycle
Optimizer	Adam	Optimizer used for gradient update
Loss Function	Cross-Entropy Loss	Loss function used for classification
Image Size	224x224	Standard input size for YOLOv8-ViT
Dropout Rate	0.3	Dropout regularization to prevent overfitting
Evaluation Metrics	Accuracy, Sensitivity, Specificity,F1-Score, MCC, AUC-ROC	Metrics to evaluate model performance
Train-Test Split	70:30:00	Split ratio for training and testing datasets
Dataset Links: MHealth : https://doi.org/10.24432/C5TW22WISDM : https://doi.org/10.24432/C5HK59		

Table 2. Matrix Analysis

Confusion Matrix	Predicted: No Activity	Predicted: Activity
Actual: No Activity	145	5
Actual: Activity	12	338

Figure 4 shows the 2x2 confusion matrix of the YOLO-ViT with the PSO & GAN model. The matrix is essential to measure the key performance metrics such as precision, recall, specificity, and F1-score. It provides insights into the model's correct predictions and error rates during every iteration. Based on this, the model is guided, tuned, and optimized to ensure the classifier achieves accuracy and maintains a balanced ratio among all classes.

2.4.1. Performance Evaluation

The performance evaluation of the proposed hybrid YOLOv8-ViT with the PSO and GAN model is initially done using MATLAB and Python over 30 iterations. The test is conducted with the combined benchmark datasets from 23,000 samples from MHealth and 1 million records from WISDM, which are pre-processed and converted into deep frame-based inputs to the

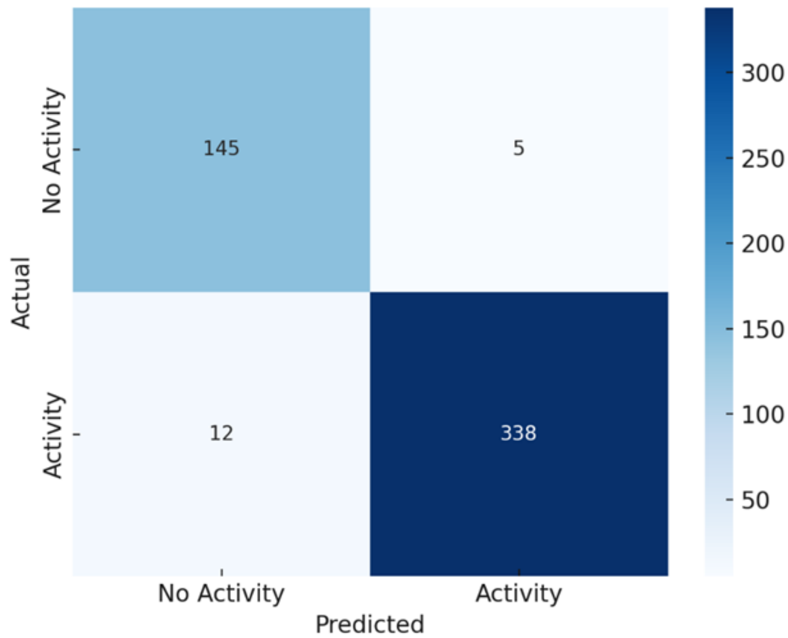

Fig 4. 2x2 Confusion Matrix for 30 Iterations

system. The data is split into a 70:30 ratio for training and testing and executed for 20 epochs per iteration with a batch size 64. Python (PyTorch) for ViT modules and GAN data augmentation, while MATLAB is used to integrate YOLOv8-like extraction of spatial features, metric evaluation, and PSO-based hyperparameter tuning. During implementation, the key parameters of MATLAB include a 0.001 initial learning rate and a 0.3 dropout rate with necessary optimizer settings. Adam optimizer is used with a custom training loop to balance the data. This dual execution environment allows the model to outperform with robust performance metrics. The results are executed using the following performance evaluation metrics formula for every iteration.

2.4.2. Performance Evaluation Metrics of YOLOv8-ViT with PSO & GAN

The following are the formula for key performance metrics used in the proposed deep learning hybrid framework. Top six potential performance metrics are utilized to showcase the model's ability in human activity recognition, classification and handling complex datasets.

$$\textbf{Accuracy} = \frac{(TPR+TNR)}{(TPR+TNR+FPR+FNR)} \times 100 \tag{9}$$

Accuracy measures the overall correctness of human activity classification across all classes in different iterations.

$$\textbf{Sensitivity} = \frac{TPR}{(TPR+FNR)} \times 100 \tag{10}$$

$$\textbf{Specificity} = \frac{TNR}{(TNR+FPR)} \times 100 \tag{11}$$

Sensitivity and Specificity predicts the TPR/Recall and TNR values and check the model's ability to detect the actual human activities in a robust manner.

$$\textbf{AUC — ROC}\,(\textbf{TP}\&\textbf{FP})\,TP = \frac{TPR}{(TPR+FNR)},\ FP = \frac{FPR}{(FPR+TNR)} \tag{12}$$

AUC-ROC provides the discriminative capability of the model across the given threshold values.

$$\textbf{F1} - \textbf{Score} = \frac{2*(Precision\ *\ Recall)}{(Precision+Recall)} \tag{13}$$

$$\mathbf{MCC} = \frac{T_1}{\sqrt{T_2 \times T_3 \times T_{4\times} T_5}} \times 100 \tag{14}$$

F1-Score balances the precision and recall to measure the robustness of the proposed model in unbalanced datasets. MCC provides a balanced evaluation for imbalanced MHealth and WISDM datasets, reflecting the high quality and accuracy of binary classification. The comparative results and discussions are shown in section 3 with table values of proposed & existing models of all metrics with a MATLAB 6 panel analysis chart.

3 Results and Discussions

This section provides a comprehensive analysis of the novel YOLOv8-ViT with PSO & GAN hybrid deep learning model for human activity recognition, which is evaluated using two major datasets, MHealth and WISDM. The performance is compared against three state-of-the-art models, such as HLA [6], SMO-DNN [7], and AMC-CNN [8], to exhibit the superiority of spatial feature extraction and human activity classification. The simulations are conducted using MATLAB and PyTorch to visualize the analysis. The quantitative findings are discussed using six performance metrics computed for both iteration 1 and iteration N (N=30). The below-mentioned tables (Tables 2 to 7) highlight the metric-wise results with 6 6-panel MATLAB chart to present the progression and comparison strengths. The X-axis in the chart denotes the model names, and the Y-axis represents the percentage. AUC-ROC curve indicates the model's discriminative capability to balance the datasets for human activity recognition to achieve generalization, classification, and robustness across different sensor readings.

3.1 Accuracy

The accuracy findings reveal the significant technical advancements of the proposed hybrid deep learning YOLOv8-ViT model integrated with PSO and GAN to boost the performance of HAR. The baseline models showed less than 85% accuracy during implementation. The proposed model reaches 95.4% accuracy, marking a substantial gain with the help of high-resolution spatial feature extraction, allowing a dynamic detection of motion characteristics from image-transformed sensor data. The ViT is enhanced by modeling long-range temporal dependencies, which enables the system to learn the sequential activity contexts. PSO finetunes the parameters, which reduces convergence time and overfitting issues. GAN does data augmentation, which increases training diversity and addresses the class imbalance. This model creates a pipeline to accelerate the learning and generalize exceptionally well across heterogeneous HAR datasets.

Table 3. Accuracy Analysis – Performance Results

Metrics / Models	HLA [6]	SMO-DNN [7]	AMC-CNN [8]	YOLOv8-ViT with PSO & GAN [Proposed]
Accuracy (IT-1)	78.6	74.5	76.1	92.8
Accuracy (IT-N)	83.2	82.6	84.7	95.4

3.2 Sensitivity & Specificity

The proposed YOLOv8-ViT model demonstrates high performance in both sensitivity and specificity, as shown in Table 4. The model surpasses all baseline algorithms limited to sub-83% ranges. High sensitivity indicates the model's reliability in recognizing true activity classes, which is essential for health monitoring and fall detection applications. YOLO captures spatial detection, such as human motion frames, while ViT minimizes the missed detections. Simultaneously, high specificity shows fewer false rates, which rejects noise, background movements, or ambiguous gestures. This is reinforced by the fact that GAN-based augmentation enables the model to learn valid and invalid activity patterns. The synergy between spatial accuracy, contextual depth, and robust optimization made it possible to perform the model with high TPR and FPR, which is suitable for deployment in sensor-rich real-time environments.

Table 4. Comparative Analysis of Packet Delivery (%)

Metrics / Models	HLA [6]	SMO-DNN [7]	AMC-CNN [8]	YOLOv8-ViT with PSO & GAN [Proposed]
Sensitivity	72.4	76.8	82.4	93.4
Specificity	73.6	74.6	80.2	96.6

3.3 AUC

The model achieved an AUC-ROC score of 0.92, where the TPR and FPR is shown in Table 5. AUC highlights the strong discriminative ability of YOLO to distinguish between activity and non-activity classes across varying thresholds. Combining YOLO's dense feature maps and ViT's attention mechanisms enhances the discrimination power in high-dimensional space. This hybrid approach dynamically focuses on salient features while preserving the temporal coherence. The PSO algorithm intelligently adjusts the activation functions and optimizer settings to smoother decision surfaces and confident predictions. Training the model with complex data is achieved by GAN to increase the robustness of the classifier across operational conditions.

Table 5. Comparative Analysis of Optimal Route Efficiency (%)

Metrics / Models	HLA [6]	SMO-DNN [7]	AMC-CNN [8]	YOLOv8-ViT with PSO & GAN [Proposed]
TPR	0.60	0.68	0.74	0.92
FPR	0.64	0.72	0.78	0.92

3.4 F1-Score

The new framework marks a significant F1-Score of 94.6%, a strong indicator of balanced performance between the precision and recall. The new YOLOv8-ViT model outperforms the existing SMO-DNN and AMC-CNN models, proving its ability to detect true positives and false positives. The YOLOv8-ViT with PSO detection system achieves the finest localization of all movements in input activity frames to minimize the misclassifications. The dropout ratio is drastically reduced, which enhances the generalization. The GAN augmentation pipeline generates the minority class samples, reinforcing the model's fairness. As F1-score balances the trade-off between actual activity and false patterns, this metric reflects the real-world readiness for deployment in HAR applications and environments. The projected F1-score validates the framework's efficacy in dense and sparse activity settings.

Table 6. Comparative Analysis of node failure rate (%)

Metrics / Models	HLA [6]	SMO-DNN [7]	AMC-CNN [8]	YOLOv8-ViT with PSO & GAN [Proposed]
F1-Score (IT-1)	74.2	78.4	84.6	91.8
F1-Score (IT-N)	72.4	76.6	78.8	94.6

3.5 Matthews Correlation Coefficient

The MCC shows a well-balanced classification performance with 93.2% even in imbalanced HAR datasets. MCC considers four confusion matrix components, TP, TN, FP, and FN, to make it more informative than accuracy and F1-score. The proposed YOLOv8-ViT with PSO & GAN significantly detects objects by isolating the motion regions with a sequence level coherence to reduce the FPR and FNR. Batch size and learning rates are tuned in response to the convergence behavior, which helps the model to stabilize across classes. To learn the boundary cases, the model utilizes data augmentation to improve the prediction consistency. Table 7 shows the MCC results, which reflect the high correlation between the actual and predicted class distributions. The substantial MCC value makes the proposed model suitable for real-world applications like healthcare and surveillance, where the class imbalance is common.

Table 7. Comparative Analysis of Latency (in ms)

Metrics / Models	HLA [6]	SMO-DNN [7]	AMC-CNN [8]	YOLOv8-ViT with PSO & GAN [Proposed]
MCC (IT-1)	68.4	74.6	80.8	91.6
MCC (IT-N)	70.2	76.4	82.8	93.2

4 Conclusion

The comprehensive study introduces an intelligent HAR framework by integrating the high-speed object detection capabilities of YOLOv8 and the contextual learning strengths of ViT with an adaptive PSO and enhanced synthetic GAN-based data

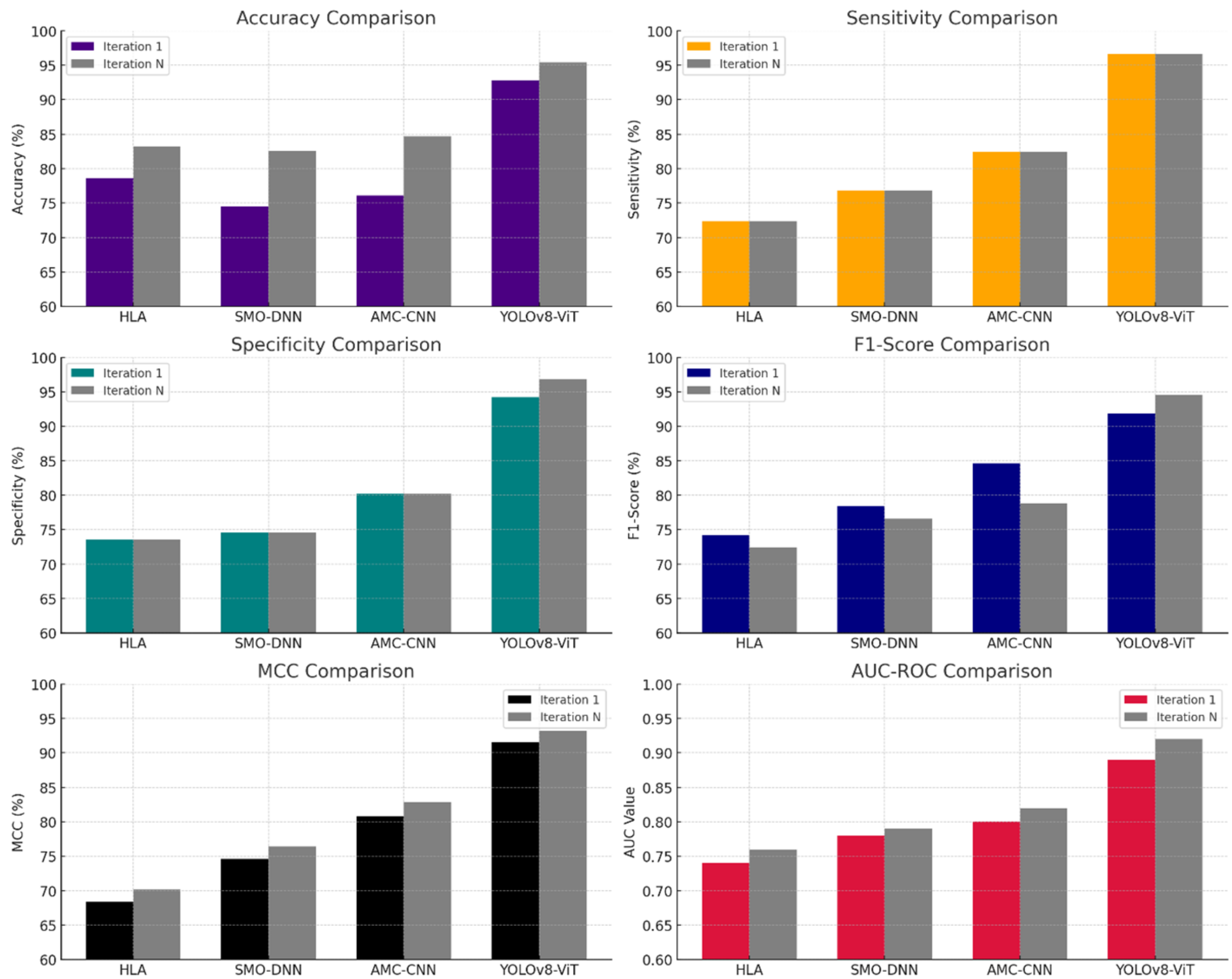

Fig 5. Six Panel Analysis Chart using MATLAB

augmentation approach, including traditional and generative approaches. Two benchmark datasets are used in this research to deliver high accuracy, robustness, and scalability of the model. The MATLAB and PYTHON experimental results confirm that the model significantly outperforms the existing conventional baseline models, such as HLA, SMO-DNN, and AMC-CNN, particularly in handling multi-sensor fusion and activity variability under complex scenarios. Achieving an accuracy of 95.4% and an AUC of 0.92, the model demonstrates higher generalization with MCC of 93.2% and F1-Score of 94.6%. These metrics highlight the performance of YOLOv8-ViT with the PSO model in real-world applications such as intelligent monitoring systems, personalized healthcare devices, and smart IoT environments. The comparative analysis shows that the proposed hybrid framework enhances human activity recognition and ensures high computational efficiency through PSO-based tuning and real-time implementation. This framework adapts data distribution shifts and user variations effectively, marking a substantial development in HAR research.

The current YOLOv8-ViT with the PSO model's limitations includes reliance on image-transformed sensor data, which leads to pre-processing overhead in critical scenarios. Additionally, using GAN may occasionally generate low variance samples (7%), which adds redundancy in the system rather than diversity. Future work may focus on integrating temporal attention mechanisms to enhance privacy and on-device adaptability.

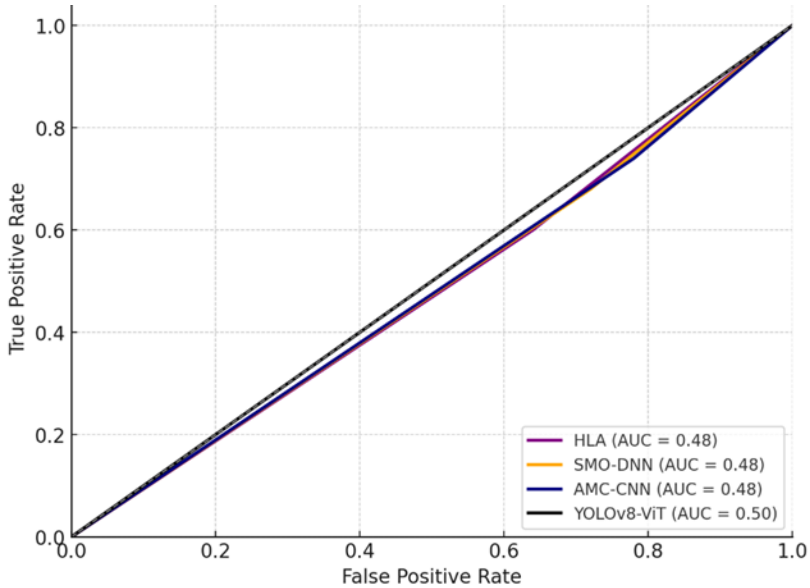

Fig 6. AUC-ROC Curve using MATLAB

List of Abbreviations

1. YOLOv8 - You Only Look Once, Version 8
2. ViT - Vision Transformer
3. GAN - Generative Adversarial Networks
4. PSO - Particle Swarm Optimization
5. GAF - Gramian Angular Fields
6. RP - Recurrence Plots
7. DLHF – Deep Learning Hybrid Framework
8. SDA – Synthesized Data Augmentation
9. AUC - Area Under the Curve of the
10. ROC - Receiver Operating Characteristic Curve

References

1) Moghaddam MG, Shirehjini AAN, Shirmohammadi S. A WiFi-Based Method for Recognizing Fine-Grained Multiple-Subject Human Activities. *IEEE Transactions on Instrumentation and Measurement*. 2023;72:1–13. https://doi.org/10.1109/tim.2023.3289547.
2) Shyam A, Ravichandran V, Preejith SP, Joseph J, Sivaprakasam M. PPGnet: Deep Network for Device Independent Heart Rate Estimation from Photoplethysmogram. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2019;p. 1899–1902. https://doi.org/10.1109/EMBC.2019.8856989.
3) Kumar A, Vishwakarma A, Bajaj V. CRCCN-Net: Automated framework for classification of colorectal tissue using histopathological images. *Biomedical Signal Processing and Control*. 2023;79(2):104172. https://doi.org/10.1016/j.bspc.2022.104172.
4) Nithyanandh S, Jaiganesh V. Interrogation of Dynamic Node Link Failure by Utilizing Bio Inspired Techniques to Enhance Quality of Service in Wireless Sensor Networks. *Shodhganga*. 2020. 10603/400513.
5) Uddin MA, Talukder MA, Uzzaman MS, Debnath C, Chanda M, Paul S, et al. Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN. *International Journal of Cognitive Computing in Engineering*. 2024;5:259–268. https://doi.org/10.1016/j.ijcce.2024.06.004.
6) Shi W, Fang X, Yang G, Huang J. Human Activity Recognition Based on Multichannel Convolutional Neural Network With Data Augmentation. *IEEE Access*. 2022;10:76596–76606. https://doi.org/10.1109/ACCESS.2022.3192452.
7) Athota RK, D S. Human Activity Recognition Based on Hybrid Learning Algorithm for Wearable Sensor Data. *Social Science Research Network (SSRN)*. 2022;p. 1–29. https://doi.org/10.2139/ssrn.4162745.
8) Kolkar R, Geetha V. Human activity recognition using deep learning techniques with spider monkey optimization. *Multimedia Tools and Applications*. 2023;82(30):47253–47270. https://doi.org/10.1007/s11042-023-15007-7.

9) Nithyanandh S, Jaiganesh V. Quality of service enabled intelligent water drop algorithm-based routing protocol for dynamic link failure detection in wireless sensor network. *Indian Journal of Science and Technology*. 2020;13(16):1641–1647. https://doi.org/10.17485/IJST/v13i16.19.
10) Sarakon S, Massagram W, Tamee K. Multisource Data Fusion Using MLP for Human Activity Recognition. *Computers, Materials and Continua*. 2025;18(2):2109–2136. https://doi.org/10.32604/cmc.2025.058906.
11) Muralidharan A, Mahfuz S. Human Activity Recognition Using Hybrid CNN-RNN Architecture. *Procedia Computer Science*. 2025;257:336–343. https://doi.org/10.1016/j.procs.2025.03.045.
12) Mennella C, Esposito M, Pietro GD, Maniscalco U. Multiscale activity recognition algorithms to improve cross-subjects performance resilience in rehabilitation monitoring systems. *Computer Methods and Programs in Biomedicine*. 2025;267:108792. https://doi.org/10.1016/j.cmpb.2025.108792.
13) Nithyanandh S, Jaiganesh V. Dynamic Link Failure Detection using Robust Virus Swarm Routing Protocol in Wireless Sensor Network. *International Journal of Recent Technology and Engineering*. 2019;8(2):1574–1578. https://doi.org/10.35940/ijrte.b2271.078219.
14) Mekruksavanich S, Phaphan W, Jitpattanakul A. A Deep Multi-Task Learning Network for Activity Recognition and User Identification Using Smartphone Sensors. *Procedia Computer Science*. 2025;256:1350–1357. https://doi.org/10.1016/j.procs.2025.02.248.
15) Trotta A, Montori F, Ciabattini L, Billi G, Bononi L, Felice MD. Edge human activity recognition using federated learning on constrained devices. *Pervasive and Mobile Computing*. 2024;104:101972. https://doi.org/10.1016/j.pmcj.2024.101972.
16) Logacjov A, Bach K. Self-supervised learning with randomized cross-sensor masked reconstruction for human activity recognition. *Engineering Applications of Artificial Intelligence*. 2024;128:107478. https://doi.org/10.1016/j.engappai.2023.107478.
17) Nithyanandh S, Jaiganesh V. Reconnaissance Artificial Bee Colony Routing Protocol to Detect Dynamic Link Failure in Wireless Sensor Network. *International Journal of Scientific & Technology Research*. 2019;10(10):3244–3251.
18) Etumusei J, Carracedo JM, McClean S. Enhancing Dynamic Human Activity Recognition Through a Novel Martingale-Based Algorithm for Change Detection. *SN COMPUTER SCIENCE*. 2025;6:451. https://doi.org/10.1007/s42979-025-03895-5.
19) Nithyanandh S, Omprakash S, Megala D, Karthikeyan MP. 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*. 2023;16(34):2753–2766. https://doi.org/10.17485/IJST/v16i34.1706.
20) Dutta SJ, Boongoen T, Zwiggelaar R. Human activity recognition: a review of deep learning-based methods. *IET Comput Vis*. 2025;p. e70003. https://doi.org/10.1049/cvi2.70003.
21) Herrmann JP, Atanasyan A, Casser F, Tackenberg S. A Petri Net Architecture for Real-Time Human Activity Recognition in Work Systems. *Procedia Computer Science*. 2023;217:1188–1199. https://doi.org/10.1016/j.procs.2022.12.317.
22) Eldho KJ, Nithyanandh S. Lung Cancer Detection and Severity Analysis with a 3D Deep Learning CNN Model Using CT-DICOM Clinical Dataset. *Indian Journal of Science and Technology*. 2024;17(10):899–910. https://doi.org/10.17485/IJST/v17i10.3085.
23) Arularasan R, Balaji D, Garugu S, Jallepalli VR, Nithyanandh S, Singaram G. Enhancing Sign Language Recognition for Hearing-Impaired Individuals Using Deep Learning. *2024 International Conference on Data Science and Network Security (ICDSNS)*. 2024;p. 1–6. https://doi.org/10.1109/ICDSNS62112.2024.10690989.
24) Selvam N, Joy JK. Plant Leaf Disease Detection with Multivariable Feature Selection Using Deep Learning AEN and Mask R-CNN in PLANT-DOC Data. *Biotech Res Asia*. 2024;21(4). https://doi.org/10.13005/bbra/3333.
25) Garcia-Gonzalez D, Rivero D, Fernandez-Blanco E, Luaces MR. New machine learning approaches for real-life human activity recognition using smartphone sensor-based data. *Knowledge Based Systems*. 2023;262:110260. https://doi.org/10.1016/j.knosys.2023.110260.