Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Efficient Skeleton-Based Human Action Recognition for IoT Edge Devices Using Lightweight LSTM Networks

Shubbar Salman Baqer Alkhafaji[1], Ali Abdulazeez Mohammed Baqer Qazzaz[2], Yousif Samer Mudhafar[3]

Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq

E-mail: [1]shubbars.alkhafaji@uokufa.edu.iq, [2]alia.qazzaz@uokufa.edu.iq, [3]yousifs.mudhafar@uokufa.edu.iq

Orcid ID: [1]0009-0009-9013-1525, [2]0000-0001-5702-0107 ,[3]0000-0003-4960-7973

Abstract

Human Action Recognition (HAR) is becoming increasingly important in areas such as intelligent surveillance, healthcare monitoring, and facilitating human-computer interaction. However, most current techniques use RGB-based deep learning, which is computationally intensive and cannot run on low-resource edge devices. This study presents a lightweight, skeleton-based Human Action Recognition (HAR) framework. The framework utilizes MediaPipe pose estimation, spatiotemporal normalization, and LSTM-based temporal modeling to efficiently recognize kinetically distinct human movements on resource-limited Internet of Things (IoT) edge devices. It combines MediaPipe pose estimation with Long Short-Term Memory (LSTM) networks to model temporal motion efficiently in resource-constrained IoT edge computing environments. Twelve kinetically unique actions from the UCF50 dataset were rendered into skeletal representations by temporally interpolating and spatially normalizing the action sequences. The HAR framework achieved 85.14% classification accuracy, indicating that sufficient motion-related information can be inferred from skeletal topology, allowing for reliable recognition at a reduced computational cost. The results demonstrate a practical balance between efficiency and performance, positioning the framework as an effective tool for edge-oriented HAR applications. From an IoT perspective, the framework has the potential to support IoT-enabled smart surveillance, assisted living, and intelligent monitoring systems by making it suitable for connected cameras and edge devices for local action recognition, even with limited computing power and communication bandwidth. This makes it a promising candidate for deployment across diverse real-world environments while preserving privacy through skeleton-based representations.

Keywords: Human Action Recognition (HAR), IoT, Skeleton-based Classification, Long Short-Term Memory (LSTM), MediaPipe, Edge Computing, Skeletal Representation.

1. Introduction

Human Action Recognition (HAR) is critical for intelligent surveillance, human-computer interaction, and fall detection systems [1]. Even though there are well-established, successful implementations of HAR using deep learning techniques like 3D Convolutional Neural Networks (CNNs) and Two-Stream models, large computational resource requirements

DOI: https://doi.org/10.36548/jtcsst.2026.3.011
Received: 13.05.2026, received in revised form: 15.06.2026, accepted: 04.07.2026, published: 16.07.2026

and concerns regarding the protection of Personally Identifiable Information (PII) hinder the deployment of these systems [2]. For example, processing high-resolution RGB video frames requires substantial computational resources (e.g., GPUs, memory bandwidth) [3], thereby making it difficult to deploy HAR systems in distributed Internet of Things (IoT) and edge computing environments where power and computational resources are limited [4]. In addition, processing raw video introduces additional PII-related challenges [5].

The rapid growth of IoT systems has increased the use of intelligent cameras and edge devices that perform local data processing without continuous cloud communication. Human Action Recognition is therefore becoming an important edge intelligence task for smart surveillance, healthcare monitoring, assisted living, and public safety applications, creating a growing demand for lightweight recognition frameworks suitable for resource-constrained IoT environments.

Traditional models that perform spatiotemporal feature extraction from RGB video streams at the pixel level [6] are very sensitive to light changes, background noise, and occlusions. They also tend to learn the context of the scene being observed rather than the actual dynamics of the motion observed [7]. Furthermore, the increased overall processing time for high-resolution video will cause increased latency issues and resource consumption, which will make deploying these solutions at the edge in real time very difficult [8].

Human Action Recognition based on skeletons works well when the discriminative characteristics of the action are primarily encoded in body kinematics rather than through the appearance of an object (i.e., "How does the object look?"). Essentially, representing the human body as a collection of joint coordinates allows a dramatic decrease in the amount of input dimension being sent while still preserving the vital amount of motion (kinematic) data necessary for recognizing actions (9). Currently, with the advent of monocular pose estimation methods, skeletal landmark data can now be captured accurately using traditional RGB cameras (10). There are many different frameworks that provide accurate, reliable, and real-time keypoint detection, including OpenPose [11] and MediaPipe [12]. (MediaPipe provides 33 3D landmarks using a compact architecture.) In addition, MediaPipe is built for use in mobile/edge computing environments, making it suitable for resource-constrained skeleton-based HAR systems (12). The temporal nature of human actions is also a critical aspect of skeleton-based recognition [13]. Although RNNs are good for modeling sequences, they often struggle to capture long-term dependencies. This difficulty is resolved through Long Short-Term Memory (LSTM) networks, which allow for long-time correlations to be encapsulated using gating mechanisms that hold temporal data in memory for long periods of time [14]. In this study, LSTM networks are employed to classify normalized skeletal coordinate sequences quickly [15].

The proposed architecture is based on a lightweight, skeleton-based Human Action Recognition (HAR) framework, which has been developed to be suitable for deployment in resource-limited environments located at the edge of the network. MediaPipe pose estimation is integrated with the LSTM networks to better accommodate the time-varying characteristics of human behavior by using skeletal representations instead of pixel-based representations during recognition. The aim of this research is to achieve a balance between recognition accuracy and computational efficiency, as opposed to achieving maximum recognition performance by using expensive, computation-intensive deep architectures. This is done by limiting recognition to those actions which have clearly defined kinematic properties and for which a skeletal representation can sufficiently and accurately represent them at the time of

recognition, thereby allowing the framework to be deployed on resource-constrained edge environments. To implement this approach, a combined processing pipeline has been created consisting of three distinct processes: (i) Extraction of skeletal topology, (ii) Spatiotemporal normalization, and (iii) Sequential temporal modeling. As a result, the proposed framework has demonstrated that reliable and accurate action recognition can be performed using a lightweight architecture, thus making the framework suitable for real-time intelligent applications. The main contributions of this study are summarized as follows:

1. A lightweight skeleton-based Human Action Recognition framework designed for deployment on resource-constrained edge devices using MediaPipe pose estimation and LSTM temporal modelling.

2. A spatiotemporal preprocessing pipeline based on temporal interpolation and hip-cantered coordinate normalization to improve the consistency of skeletal motion representations.

3. An experimental evaluation using a kinetically distinct subset of the UCF50 dataset to investigate lightweight skeleton-based recognition for actions whose primary discriminative features are encoded in body motion rather than object appearance.

4. An analysis of the trade-off between computational efficiency, skeletal-coordinate representation, and recognition performance for edge-oriented Human Action Recognition systems, with particular emphasis on deployment scenarios involving IoT-enabled smart sensing environments.

2. Related Works

Due to their ability to gather information about forward motion and appearance (spatial), recent advances in Human Action Recognition (HAR) have been largely driven by deep learning models operating on RGB video sequences. For example, Xiong et al. [16] proposed a two-stream dilated 3D neural network that combines RGB and skeletal information to increase recognition accuracy by considering complementary spatial and temporal cues together. In addition, Liu et al. [17] introduced a cross-scale cascade multimodal fusion transformer in the RGB-D action recognition task by showing that multimodal feature fusion using transformer architectures can produce superior performances in large benchmark datasets. Despite the high accuracy obtained from previous methods, the amount of computation and high dimensionality of the visual features used pose limitations when deploying on resource-constrained edge computing devices.

To reduce computational complexity, increasing attention has been directed toward skeleton-based Human Action Recognition (HAR) where each human motion is represented using the coordinates (or locations) of the key joints in their bodies (skeletons) instead of the pixel data from images. Doan [18] combined MediaPipe pose estimation with an LSTM network in order to develop a very fast and efficient patient activity recognition system capable of real-time execution on edge devices. Song et al. [19] produced an improved skeleton-based recognition system using the EfficientGCN architecture that takes advantage of optimized graph convolutions and spatial-temporal attention mechanisms to achieve high recognition accuracy while reducing computational cost compared with conventional graph models. Zhong et al. [20] presented an architecture for performing hand gesture recognition in real-time using

hybrid convolution (CNN)-transformer architectures working directly on skeleton coordinate systems. Wang et al. [21] utilized MoveNet with ConvLSTM and attention mechanisms to recognize badminton actions under multiple viewpoints, achieving excellent recognition performance at the expense of increased architectural complexity.

Aside from the aforementioned work, various studies focused on lightweight Human Action Recognition (HAR) systems for edge computing or IoT deployment have also been published. As such, the number of studies published concerning the creation of lightweight HAR systems has focused on developing lightweight HAR systems with multiple studies documenting techniques and/or frameworks for developing a lightweight HAR system. One such study by Wazwaz et al. [22] discusses the development of an innovative Raspberry Pi based HAR framework that utilizes smartphone sensor data and incorporates a LightGBM classification model to yield exceptionally accurate results with minimal latency. Alotaibi et al. [1] presented similar findings utilizing deep neural networks for smart home change detection with ambient sensor data; however, they achieved great success in recognizing activities without any reliance on visual or camera data. Although these approaches are computationally efficient, they rely on wearable or ambient sensing technologies rather than camera-based skeletal motion representations.

To this point, researchers have demonstrated significant progress in Human Action Recognition (HAR) through the application of RGB-based deep learning models, graph convolutional networks, transformer architectures, and skeleton-based temporal models. Nevertheless, many high-performing approaches rely on computationally intensive architectures, multimodal sensing, or application-specific datasets that limit their suitability for resource-constrained edge deployment. Importantly, existing lightweight skeleton-based methods have primarily focused on healthcare monitoring or other application-specific applications and relatively little attention has been given to evaluating skeleton-only action recognition on publicly available benchmark datasets under edge-computing constraints. Motivated by this gap, the present study investigates a lightweight framework that combines MediaPipe pose estimation, spatiotemporal normalization, and LSTM-based temporal modelling to evaluate the feasibility of efficient skeleton-based Human Action Recognition for resource-constrained IoT edge computing environments while reducing dependence on pixel-level visual representations during the recognition stage.

Table 1. Summary of the related work

Ref.	Dataset	Methodology	Strengths	Limitations
[1]	ARAS Smart Home Dataset	ANN, CNN, and RNN architectures	High recognition accuracy for ambient sensor-based activity recognition without visual sensing.	Relies on environmental sensors rather than skeletal or vision-based motion analysis.
[23]	FPHA, DHG14/28	2D Deep Video Capsule Network with Temporal Shift Module (TSM)	Efficient temporal modeling with significantly fewer parameters than conventional 3D CNNs while maintaining competitive recognition performance.	Evaluated mainly on hand gesture datasets and requires relatively complex capsule-based training.
[18]	KARD + Custom Patient Dataset	MediaPipe Pose + Multi-layer LSTM	Lightweight edge-oriented framework capable of real-time inference with high recognition accuracy using skeletal landmarks.	Evaluation is limited to healthcare monitoring activities with relatively small action diversity.
[19]	NTU RGB+D 60/120	EfficientGCN with Spatial-	High recognition accuracy through optimized graph-based skeletal	Graph-based architecture remains more computationally

		Temporal Joint Attention	modeling with improved computational efficiency.	demanding than lightweight recurrent models.
[16]	UCF101, HMDB51	Two-Stream Dilated 3D CNN (RGB + Skeleton)	Combines appearance and motion information to improve action recognition robustness.	High computational complexity and reduced suitability for real-time edge deployment.
[22]	Smartphone Accelerometer Dataset	PCA + LightGBM on Raspberry Pi	Extremely fast inference with low computational requirements for IoT activity recognition.	Limited to wearable sensor data and does not support vision-based action recognition.
[17]	NTU RGB+D, PKU-MMD	Cross-Scale Cascade Multimodal Fusion Transformer	Strong multimodal feature fusion achieving high recognition performance for RGB-D action recognition.	Transformer architecture requires substantial computational resources unsuitable for lightweight edge systems.
[20]	Briareo and Multimodal Hand Gesture Datasets	Hybrid 3D-CNN + Transformer on Skeleton Coordinates	Real-time hand gesture recognition with efficient temporal modeling on CPU platforms.	Focused exclusively on hand gesture recognition rather than full-body human actions.
[21]	Badminton Action Dataset	MoveNet + ConvLSTM + SE Attention	Effective modeling of complex skeletal motion with strong recognition performance under multiple viewpoints.	Higher computational complexity than conventional LSTM architectures and limited to sports-specific actions.
[24]	UCF101	Deep LSTM with IIWBPR Optimization and Handcrafted Features	Strong temporal modeling with improved convergence on large-scale action recognition datasets.	Depends on handcrafted feature extraction, increasing preprocessing complexity and reducing suitability for lightweight edge deployment.

Table 1 shows that current Human Action Recognition systems improve action recognition performance primarily due to the use of graph neural networks and transformer architectures, multimodal fusion techniques, and specific datasets. Unfortunately, most of these systems increase the complexity of computations or require a specialized sensor type, which makes them unsuitable for deployment in resource-constrained environments. Motivated by these observations, the present study investigates a lightweight skeleton-based framework that emphasizes computational efficiency and practical edge deployment while maintaining effective action recognition performance.

3. Methodology

The proposed framework introduces a lightweight, computation-efficient pipeline for Human Action Recognition (HAR) by transforming high-dimensional video data into lightweight skeletal topology sequences that are processed by a sequential recurrent model. The overall design of this system, depicted in Figure 1, consists of four stages:

- Dataset curation,
- Pose estimation and feature extraction
- Spatiotemporal normalization
- Sequential modeling using Long Short-Term Memory (LSTM) networks.

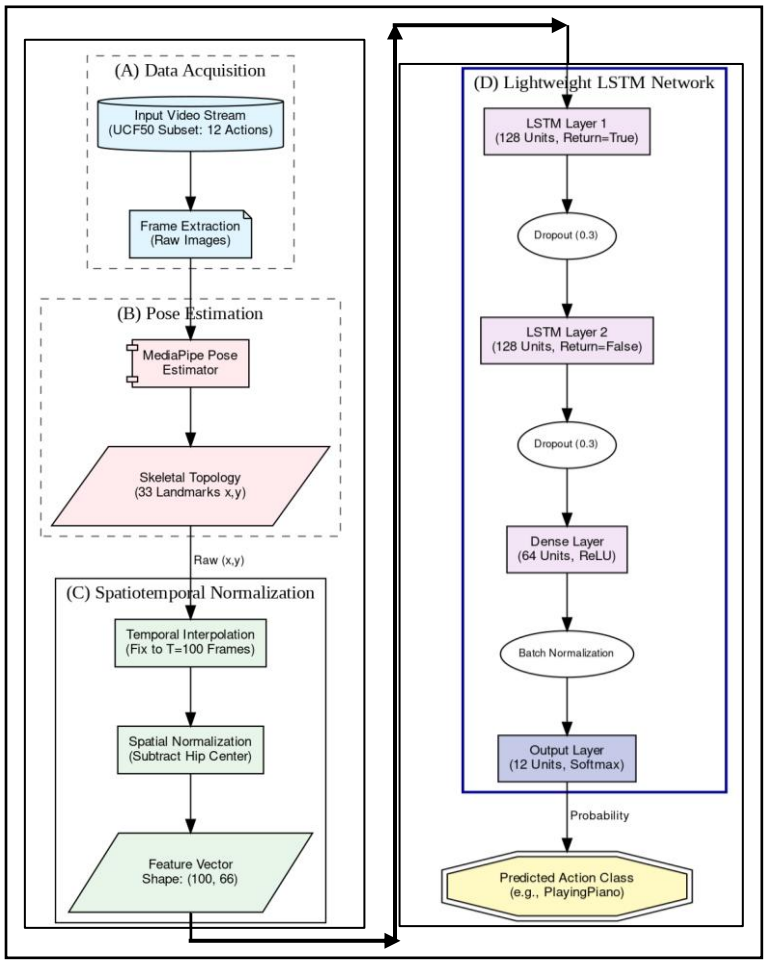

Figure 1. Workflow diagram of the proposed system.

3.1 Dataset Selection and Construction

The UCF50 dataset contains 50 action categories and provides a wide range of human activities that vary in their motion characteristics. To evaluate the lightweight skeletal based recognition framework, 12 classes of actions were selected from the UCF50 dataset. To ensure dataset transparency and address class balance, Table 2 reports the number of samples per action class used in this study.

Table 2. Distribution of action samples per Class

#	Action Class	Number of Samples
1	JumpingJack	60
2	PushUps	140
3	PullUps	124
4	GolfSwing	140
5	PlayingPiano	100
6	PlayingViolin	100
7	RockClimbingIndoor	116
8	SkateBoarding	148
9	Lunges	108
10	YoYo	136
11	Punch	100
12	Diving	128
	Total	1400

The twelve action classes included in the framework were selected based on three criteria. First, the selected actions must consist of a wide variety of upper body, lower body and full body movements. Second, the selected actions must display unique kinematic patterns of motion correlated to the skeletal system permitting analysis based on the skeletal structure. Third, actions that mainly depend on object manipulation, environmental context, or appearance were excluded because skeletal topology alone cannot adequately represent such information. Jumping Jacks, Push Ups, Pull Ups, Golf Swing, Playing Piano, Playing Violin, Indoor Rock Climbing, Skateboarding, Lunges, Yo-Yo, Punch and Diving were the selected actions; The objective was not to maximize benchmark accuracy but to evaluate whether a lightweight skeleton-only framework can effectively recognize kinetically distinct actions whose primary discriminative information is contained in body motion rather than object appearance under practical IoT edge-computing constraints.

While the study identified 12 action classes to evaluate the proposed lightweight skeleton-based framework under conditions where body kinematics serve as the primary source of discriminative information, it also emphasizes the evaluation of the feasibility of efficient skeleton-only recognition for resource-constrained IoT edge devices, in contrast to maximizing benchmark performance among all available action classes across the UCF50 dataset. As the evaluation excluded actions that rely primarily on manipulated objects and a rich environment context, results presented here should not be interpreted as representing the full diversity of the UCF50 dataset. Future work will expand the framework to include additional benchmark datasets and action categories to further validate the framework's generalization capability.

3.2 Topological Feature Extraction (Pose Estimation)

The MediaPipe Pose framework extracts skeletal landmarks of very high fidelity from each RGB Frame. After obtaining landmarks, the subsequent recognition stage operates exclusively on geometric skeletal coordinates, rather than the pixel level features of the image. Therefore, the temporal classifier processes only compact joint representations, substantially reducing the visual information required while preserving the temporal characteristics of human motion. Let represents the skeleton vector at time frame t. The feature extraction process maps the raw image to a coordinate vector:

$$L_t = \{(x_i, y_i)_t \mid i = 1, 2, 3, \dots, 33\} \quad (1)$$

Where (x_i, y_i) denote the normalized screen coordinates of the i-th landmark (e.g., nose, left shoulder, right knee). Consequently, only normalized two-dimensional landmark coordinates are retained, the input data's dimensions are decreased from 99 feature sets (i.e., 33 x 3) to 66 features (i.e., 33 x 2) per frame while maintaining sufficient skeletal motion information for temporal action recognition. Although MediaPipe Pose estimates three-dimensional landmark coordinates (x, y, and z), only two-dimensional (x, y) normalized landmark positions are used in this work for the purpose of developing a lightweight framework for resource-constrained IoT edge devices. The z-coordinate estimated from a monocular RGB camera represents relative depth rather than true metric depth and is generally more susceptible to viewpoint variations and estimation noise. Using only the two-dimensional skeletal landmark positions reduces the input dimensionality from 99 to 66 features per frame, thereby reducing computational complexity while providing sufficient skeletal motion data to recognize the distinct characteristics of the selected kinetically distinct human actions.

3.3 Spatiotemporal Normalization

The original skeletal data has a significant amount of noise generated by the relative motion of the camera, the physical size of the human subject being filmed, and the speed at which the video was filmed. To enable the model to learn about the action dynamics rather than the noise in the environment, an extensive preprocessing pipeline is applied to remove this noise before training occurs.

3.4 Temporal Interpolation (Sequence Standardization)

Video clips in the UCF50 dataset vary in duration. To generate a fixed-size input tensor for the LSTM network, linear interpolation is applied to normalize all sequences to a fixed length of T=100 frames. For a video with N original frames, the new sequence is generated by linearly interpolating the original skeletal landmarks over a uniform temporal grid. The temporal interpolation process is mathematically formulated as follows:

$$t_i = \frac{i(N-1)}{T-1}, \quad i = 0,1,\dots,T-1 \tag{2}$$

where N is the number of frames in the original sequence, T is the target sequence length (100 frames), and t_i denotes the temporal position corresponding to the i-th frame in the normalized sequence. The interpolated skeletal coordinate vector is then computed using linear interpolation:

$$S'^{(i)} = (1-\alpha)S(\lfloor t_i \rfloor) + \alpha S(\lceil t_i \rceil) \tag{3}$$

Where $\alpha = t_i - \lfloor t_i \rfloor$ and $S(\cdot)$ represents the original skeletal coordinate vector, while $S'(i)$denotes the interpolated skeletal coordinate vector at the normalized frame. This interpolation preserves the temporal ordering of the skeletal motion while producing fixed-length sequences suitable for LSTM training. This ensures that fast and slow executions of the same action are mapped to a consistent temporal dimension.

3.5 Spatial Normalization (Translation Invariance)

To make the recognition system invariant to the subject's position within the camera frame, all landmark coordinates are transformed relative to the body's center of gravity. The Hip Center (C_{hip}) is calculated as the midpoint between the left hip (p_{23}) and right hip (p_{24}):

$$C_{hip} = \frac{p_{23} + p_{24}}{2} \tag{4}$$

Subsequently, this center point is subtracted from all 33 landmarks in the frame:

$$\acute{P}_i = P_i - C_{hip} \;\; \forall\, i \in \{1,\dots,33\} \tag{5}$$

This translation ensures that the coordinate (0, 0) always represents the center of the subject's hips, forcing the neural network to focus solely on the relative movement of limbs.

3.6 Lightweight LSTM Network Architecture

The proposed network receives normalized skeleton sequences of size (100, 66) and processes them using two LSTM layers with 128 hidden units, separated by dropout layers (0.3). A fully connected ReLU layer and batch normalization are then applied before a Softmax output layer that predicts the twelve action classes. The complete model contains 240,716 trainable parameters while maintaining a lightweight architecture suitable for edge and IoT applications. The architecture was intentionally designed to balance temporal modelling capability with computational efficiency for deployment on resource-constrained IoT edge devices. The architecture of the proposed network is summarized in Table 3.

Table 3. Model summary

Layer Type	Output Shape	Parameters	Function
Input	(None, 100, 66)	0	Receives normalized skeleton sequence
LSTM	(None, 100, 128)	99,840	Extracts temporal sequence features
Dropout	(None, 100, 128)	0	Regularization (0.3)
LSTM	(None, 128)	131,584	Compresses temporal dynamics
Dropout	(None, 128)	0	Regularization (0.3)
Dense (ReLU)	(None, 64)	8,256	High-level feature interpretation
BatchNormalization	(None, 64)	256	Stabilizes weights
Dense (Softmax)	(None, 12)	780	Final Classification
Total Params	240,716	240,716	Lightweight Architecture

The proposed architecture contains only 240,716 trainable parameters, making it substantially smaller than many graphs neural network and transformer-based HAR architectures reported in the literature, thereby supporting efficient deployment on resource-constrained IoT edge devices. The selected hyperparameters were chosen based on preliminary experiments to achieve an optimal balance between recognition accuracy and computational efficiency. A two-layer LSTM with 128 hidden units provided sufficient capacity to capture temporal dependencies in the skeletal sequences. A dropout rate of 0.3 effectively reduced overfitting without degrading model performance. This configuration was used in all experiments.

3.7 Implementation and Training Configuration

The framework was implemented using Python 3.10 and TensorFlow/Keras. Model training employed a stratified 75%/25% train-test split, the Adam optimizer, categorical cross-entropy loss, a batch size of 32, and a maximum of 120 epochs. Early stopping with a patience of 15 epochs was applied to improve generalization and reduce unnecessary computation.

4. Results and discussion

This section evaluates the model performance through comprehensive metrics, experimental computational validation, and quantitative comparisons with state-of-the-art methods.

4.1 Performance Analysis of the Proposed System

The developed lightweight skeleton-based framework achieved an overall classification accuracy of 85.14%, showing that temporal modeling of skeletal landmarks can provide enough discriminative information to classify human actions without depending on pixel-level

image features for recognition. The stable convergence achieved in the training process indicates that the combination of temporal interpolation, hip-centered spatial normalization, and LSTM-based sequence modeling generates consistent skeletal representations, which can be used for classifying actions based on these representations. The class-wise results indicate that recognition performance is strongly influenced by the uniqueness of the underlying body kinematics. Actions characterized by large-amplitude and highly distinctive body movements achieved much higher recognition performance (i.e., Jumping Jack and Push-Ups) since their skeletal trajectories have very definitive temporal patterning that are easily learned through RNNs. Conversely, actions involving greater viewpoint variation or partial body occlusion (e.g., Diving and Indoor Rock Climbing) achieved lower recognition accuracy because pose estimation errors propagate through the temporal classification process. While overall accuracy is one way to evaluate the performance of a model, Table 4 contains classification metrics, including precision, recall, and F1-score, to provide the reader with a more comprehensive view of how well the model performs. A confusion matrix was used to analyze inter-class misclassifications.

Table 4. Comprehensive classification metrics

Metric	Score (%)
Overall Accuracy	85.14%
Macro Average Precision	86.00%
Macro Average Recall	86.00%
Macro Average F1-Score	85.00%

The confusion matrix obtained from the test set (350 samples) is displayed in Figure 2. The diagonal elements of the matrix indicate the total number of samples correctly classified for each individual class, while the off-diagonal elements are samples incorrectly classified between different classes.

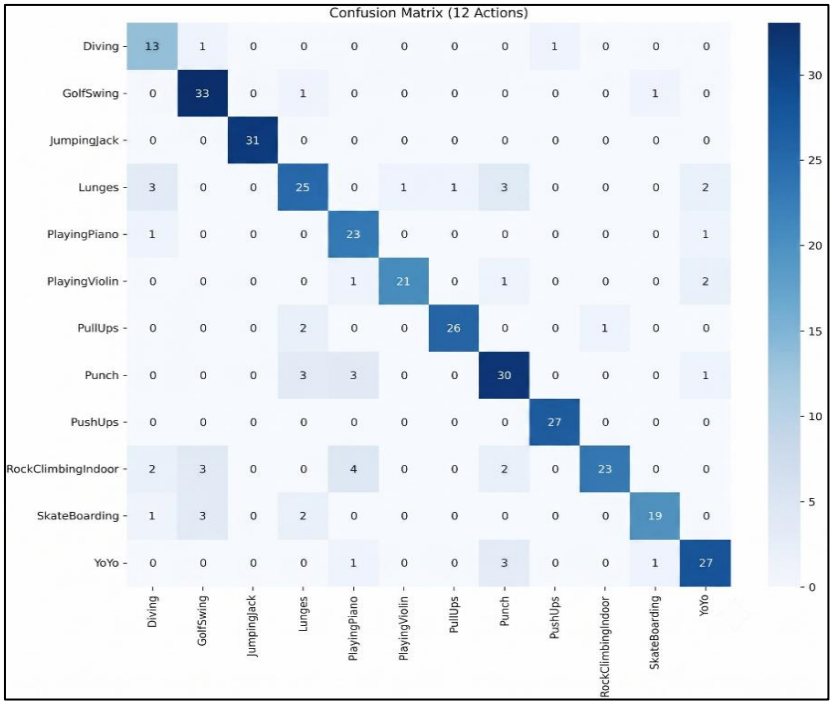

Figure 2. Confusion matrix of the proposed HAR model on the test set.

The framework also demonstrated strong recognition capability for fine-grained activities such as Playing Piano and Playing Violin despite only using skeletal information. It

did not include any pixel-level object appearance and relied only on the relative movement of joints in upper-body coordination. This suggests that temporal coordination between upper-body joints contains sufficient motion information to differentiate between certain fine-grained activities even in the absence of visually contextual clues. Overall, the experimental results highlight an important trade-off in skeleton-based Human Action Recognition where there is strong dependency on an object's appearance. However, while the elimination of pixel-level appearance would reduce the ability to recognize strong object-dependent activities, compact skeletal representation significantly reduces the input dimension and computational burden, making it attractive for deployment on resource-constrained edge devices where lightweight inference is prioritized over maximum benchmark accuracy.

4.2 Comparison with Existing Skeleton-Based and RGB-Based HAR Approaches

The framework discussed in this paper should be interpreted within the context of lightweight skeleton-based Human Action Recognition systems, and it should not be compared to high-complexity RGB-based deep learning models. Several studies indicate that skeleton representations provide a practical balance between recognition performance and computational efficiency by using body joint locations instead of pixels in an image to show humans moving. In one such study by Doan [18], MediaPipe pose estimation has been combined with an LSTM Network to develop an edge device application for recognising patient activity with the Raspberry Pi 4 [25], with success seen through its use in a healthcare monitoring application. Similarly, Wang et al. [21] combined MoveNet with ConvLSTM and attention mechanisms to recognize badminton actions, achieving high recognition performance but at the potential cost of added architectural complexity. To strengthen the contribution of this work, Table 5 provides a quantitative performance comparison between the proposed model and recent state-of-the-art HAR methods.

Table 5. Quantitative comparison with recent HAR models

Reference	Dataset	Method	Accuracy (%)	Model Size / Complexity	Edge Suitable
Xiong et al. [16]	UCF101, HMDB51	Two-Stream 3D CNN	95.6	High	No
Jandhyam et al. [24]	UCF101	Deep LSTM + IIWBPR	92.3	Medium	Partial
Proposed	UCF50 (12 classes)	MediaPipe + Lightweight LSTM	85.14	0.94 MB	Yes

As can be seen from Table 5, there are differences in recognition accuracy that should be interpreted considering deployment requirements and resource limitations. While the methods in [16] and [24] attain the highest accuracy via the use of computationally intense architectures, our proposed method provides an alternative with a strong emphasis on the computational efficiency of resources within IoT environments. Even at a lower achieving accuracy of 85.14%, our method allows for reliable human action recognition while maintaining a compact model of 0.94 MB and an inference latency of 21.40 ms; indicating the trade-offs can be made between the performance of human action recognition and the computational requirements to deliver this level of recognition performance, making this proposed method a viable candidate for use within lightweight IoT edge applications.

The present study differs from these approaches, in that it will explore the application of a lightweight MediaPipe-LSTM framework as evaluated using a subset of the publicly available UCF50 benchmark dataset under the constraints of an edge-computing environment. The goal is not to surpass benchmark accuracy provided by conventional state-of-the-art deep

learning methods, but to determine the potential of combining compact skeletal features and lightweight temporal models as an efficient solution for Human Action Recognition on resource-constrained devices. Action recognition is established in the literature, but because the datasets, action categories, evaluation criteria, and experimental conditions used by each study vary widely, the numerical accuracy comparisons provided by these studies should be used with caution when directly compared numerically. The framework proposed in this paper shows the practical trade-off that exists between computational simplicity and recognition capability for kinetically distinct actions that can be effectively represented by skeletal motion, making this model suitable for resource-constrained IoT edge applications.

4.3 Computational Complexity Analysis

The model's suitability for deployment on edge devices was experimentally validated through the measurement of computational metrics, as opposed to being merely claimed. This model achieves an average inference latency of 21.40 ms per sample, supporting real-time processing requirements (>30 FPS) and has a total model size of 0.94 MB. Therefore, these measurements indicate the model’s suitability for use on low-resource devices in a real-time application and serve to establish its feasible for IoT edge environments.

4.4 Ablation Study

A rigorous ablation study was conducted to analyze how the number and type of action classes affect the LSTM's learning capability. Two distinct experimental setups derived from the development phase were compared.

Case 1: Preliminary Micro-Experiment (4 Actions)

In the initial proof-of-concept phase, the LSTM was trained on a small subset of 4 distinct actions: [Punch, PushUps, PullUps, and Swing].

- **Observation:** The model achieved extremely high accuracy (>95%) very quickly.
- **Analysis:** With only 4 classes, the "decision boundaries" in the feature space are wide. The LSTM easily distinguished between the vertical motion of PullUps and the horizontal orientation of PushUps.
- **Limitations:** While promising, this setup is too limited for real-world applications, as it does not reflect the complexity of daily human activities.

Case 2: The "Scale-Up" Stress Test (20 Actions)

Then, the dataset was expanded to 20 actions, introducing complex classes such as Biking, HorseRiding, and Skiing.

- **Result:** The accuracy dropped significantly to 61.66%.
- **Justification:** A primary cause of failure for skeleton-only systems was object dependency. The riding of horses, bicycles, etc., was tied to the object being used, not just how the rider was positioned. By removing the visual context of the object being ridden (bike, horse), the skeleton was rendered in a neutral posture (sitting) and would have caused high amounts of confusion associated with all

other types of sitting activities. It became evident that simply adding classes without filtering for kinetic distinctiveness, resulted in degraded performance.

Case 3: Impact of Hyperparameter Tuning (Optimization)

Finally, the impact of model hyperparameters was investigated. The work compared the configuration used in the 20-action failure against the optimized 12-action success. The impact of the selected hyperparameters is summarized in Table 6.

Table 6. Impact of hyperparameter tuning

Parameter	Initial Config (20 Actions)	Optimized Config (12 Actions)	Impact on Performance
Dropout Rate	0.4 / 0.5	0.3 / 0.3	Reducing dropout prevented underfitting on the smaller feature vectors.
Epochs	150	120 (with Early Stopping)	Preventing over-training saved time and reduced overfitting.
Class Selection	Random (Object-based)	Kinetic-based	Removing object-dependent classes boosted accuracy from ~61% to 85.14%.

The ablation study indicates that lightweight skeleton-based recognition performs most effectively when actions are primarily distinguished by body kinematics rather than external object interactions, while appropriate regularization further improves temporal feature learning.

4.5 Practical Advantages of the Proposed Framework

The proposed framework has limitations. Since it is based on skeletal landmarks, it is challenging to determine actions that rely heavily upon manipulated objects or the surrounding environment. The accuracy of the recognition process also depends on pose estimation quality; therefore, recognition accuracy may decrease under occlusion or challenging viewpoints. Additionally, only using monocular RGB input will limit the motion information related to depth. Future work may investigate multimodal sensing as well as exploring graph-based skeletal modelling techniques, while ensuring they can also work within the constraints of the edge networked/IoT deployment model. The proposed framework has shown good performance in recognizing kinetically different actions, but its ability to generalize to object-centric or context-determined actions is still limited since skeletal coordinates alone cannot fully represent object appearance and scene characteristics.

4.6 Limitations and Future Work

The proposed framework combines skeletal representations with a lightweight LSTM architecture that allows for efficient Human Action Recognition in resource-limited environments. By encoding human activity in skeletal joint coordinates, rather than in pixel-level image features, the input complexity is reduced, but this data retains the time series motion data necessary for action classification. The lightweight design supports deployment on embedded processors, IoT edge devices, and network-connected cameras where local inference and low computational overhead are essential. Future studies will look into larger benchmark datasets containing a wider variety of action types and multiple modes of representation to further enhance the generalizability of the framework, and to research multimodal fusion, graph-based skeletal modeling, robustness to pose estimation error, and ways to accelerate the implementation of this framework on FPGA platforms [26] or other

embedded AI platforms with a goal of maintaining the processing efficiency needed for IoT edge deployment.

5. Conclusion

In this research, we presented a lightweight, skeleton-based Human Action Recognition framework for resource-constrained IoT edge computing environments. The combination of the MediaPipe Pose Estimation and Long Short-Term Memory models yielded an 85.14% recognition rate on the kinetically distinct subset of the UCF50 dataset. The results indicate that skeletal models can be used to effectively represent actions and facilitate an efficient processing pipeline for edge computing applications. Recognition performance was highest for actions with distinctive skeletal motion patterns and was negatively impacted when the actions had contextual (environmental) cues, were viewed at varied angles (viewpoint), and/or were occluded. In summary, the results show how lightweight, skeleton-based recognition systems can optimally utilize the trade-offs between computational efficiency and contextual information. Although the proposed framework is not intended to compete with computationally intensive multimodal architectures, it can provide a viable alternative for IoT edge intelligence applications that require lightweight inference and resource efficiencies. In summary, the findings indicate that skeletal-coordinate representations provide a practical balance between recognition performance and computational simplicity within geographically distributed IoT sensing environments. Therefore, the proposed framework may be a suitable candidate for intelligent IoT sensing systems where lightweight local action recognition must occur using only minimal computational resources. The proposed framework is particularly suitable for applications in which body kinematics provide the dominant source of discriminative information, while object appearance plays a secondary role, making it well-suited for lightweight IoT edge intelligence systems.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

[1] Alotaibi, Faiz, Mrim M. Alnfiai, Fahd N. Al-Wesabi, Mesfer Alduhayyem, Anwer Mustafa Hilal, and Manar Ahmed Hamza. "Optimal Deep Recurrent Neural Networks for Iot-Enabled Human Activity Recognition in Elderly and Disabled Persons." Journal of Disability Research 2, no. 2 (2023): 79-88.

[2] Martin, Pierre-Etienne, Jenny Benois-Pineau, Renaud Péteri, Akka Zemmari, and Julien Morlier. "3D Convolutional Networks for Action Recognition: Application to Sport Gesture Recognition." In Multi-Faceted Deep Learning: Models and Data, Cham: Springer International Publishing, 2012, 199-229.

[3] Eleeasa, Kevina Nugraha, Salomo Hendrian Sudjonoa, and Alexander Agung Santoso Gunawanb. "Video Based Physical Bullying Detection with Deep Learning: Utilizing 3 Dimensional and Dynamic Vision Sensors." Procedia Computer Science 269 (2025): 1077-1086.

[4] Peng, Zhiyong, Jiang Du, and Yulong Qiao. "Design of GPU Network-On-Chip for Real-Time Video Super-Resolution Reconstruction." Micromachines 14, no. 5 (2023): 1055.

[5] Spiller, E., & Giampaolo, N. (2022). Social, ethical, legal, privacy issues identification and monitoring (HAIKU Project Deliverable D1.3).

[6] Wang, Cailing, and Jingjing Yan. "A Comprehensive Survey of Rgb-Based and Skeleton-Based Human Action Recognition." IEEE Access 11 (2023): 53880-53898.

[7] Khedher, Issam, Jean-Marie Favreau, Serge Miguet, and Gilles Gesquière. "RGB2LST: Enhancing Deep Learning-Based Land Surface Temperature Estimation with Multi-Modality and Artifacts Removal." In 2024 32nd European Signal Processing Conference (EUSIPCO), IEEE, 2024, 2002-2006.

[8] O'Donnell, Kyle, and Chandra Kambhamettu. "Feature Matching in the Dark: Homography-Based RGB-IR Feature Transformation for Low-Light Vision." In Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, 1703-1711.

[9] Yang, Haoxin, Weihong Chen, Xuemiao Xu, Cheng Xu, Peng Xiao, Cuifeng Sun, Shaoyu Huang, and Shengfeng He. "Starpose: 3D Human Pose Estimation via Spatial-Temporal Autoregressive Diffusion." IEEE Transactions on Circuits and Systems for Video Technology (2025).

[10] Pan, Xiaojian, Guang Li, Ningfei Zhang, and Jianjun Li. "3D Human Pose Estimation Based on a Hybrid Approach of Transformer and GCN-Former." Journal of Visual Communication and Image Representation (2025): 104696.

[11] Jiang, Jia-Hong, and Nan Xia. "Pcnet: A Human Pose Compensation Network Based on Incremental Learning for Sports Actions Estimation." Complex & Intelligent Systems 11, no. 1 (2025): 17.

[12] Chen, Kaixin, Lin Zhang, Zhong Wang, Shengjie Zhao, and Yicong Zhou. "Skeleton-Aware Graph-Based Adversarial Networks for Human Pose Estimation from Sparse Imus." ACM Transactions on Multimedia Computing, Communications and Applications 21, no. 4 (2025): 1-22.

[13] Ali, Khalid, Zineddine Bettouche, Andreas Kassler, and Andreas Fischer. "Enhancing Spatiotemporal Networks with Xlstm: A Scalar LSTM Approach for Cellular Traffic Forecasting." In 2025 16th International Conference on Network of the Future (NoF), IEEE, 2025, 167-175.

[14] Selvaperumal, P., F. Sheeja Mary, T. L. Roy, Yousef A. Baker El-Ebiary, Gowrisankar Kalakoti, and Sandeep Kumar Mathariya. "Explainable Deep Temporal Modeling for Stroke Risk Assessment Using Attention-Based LSTM Networks." International Journal of Advanced Computer Science & Applications 16, no. 6 (2025).

[15] Chakri, Sana, and Naoual Mouhni. "Time-Series Neural Network Predictions Using LSTM for Clustering and Forecasting GPS Data of Bird Immigration." In 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST), IEEE, 2024, 1-4.

[16] Xiong, Xin, Weidong Min, Qing Han, Qi Wang, and Cheng Zha. "Action Recognition Using Action Sequences Optimization and Two-Stream 3D Dilated Neural Network." Computational Intelligence and Neuroscience 2022, no. 1 (2022): 6608448.

[17] Liu, Zhen, Qin Cheng, Chengqun Song, and Jun Cheng. "Cross-Scale Cascade Transformer for Multimodal Human Action Recognition." Pattern Recognition Letters 168 (2023): 17-23.

[18] Doan, Thanh-Nghi. "An Efficient Patient Activity Recognition Using LSTM Network and High-Fidelity Body Pose Tracking." International Journal of Advanced Computer Science and Applications 13, no. 8 (2022). 226–232.

[19] Song, Yi-Fan, Zhang Zhang, Caifeng Shan, and Liang Wang. "Constructing Stronger and Faster Baselines For Skeleton-Based Action Recognition." IEEE transactions on pattern analysis and machine intelligence 45, no. 2 (2022): 1474-1488.

[20] Zhong, Enmin, Carlos R. Del-Blanco, Daniel Berjón, Fernando Jaureguizar, and Narciso García. "Real-Time Monocular Skeleton-Based Hand Gesture Recognition Using 3D-Jointsformer." Sensors 23, no. 16 (2023): 7066.

[21] Wang, Chuanchuan, Ahmad Sufril Azlan Mohmamed, Mohd Halim Bin Mohd Noor, Xiao Yang, Feifan Yi, and Xiang Li. "ARN-LSTM: A Multi-Stream Fusion Model for Skeleton-Based Action Recognition." arXiv preprint arXiv:2411.01769 (2024).

[22] Wazwaz, Ayman, Khalid Amin, Noura Semary, and Tamer Ghanem. "Light Weight Human Activity Recognition using Raspberry PI IoT Edge and Reduced Features from Smartphones." IJCI. International Journal of Computers and Information 10, no. 3 (2023): 1-8.

[23] Voillemin, Theo, Hazem Wannous, and Jean-Philippe Vandeborre. "2D Deep Video Capsule Network with Temporal Shift for Action Recognition." In 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, 3513-3519.

[24] Jandhyam, Lakshmi Alekhya, Ragupathy Rengaswamy, and Narayana Satyala. "An Optimized Deep LSTM Model for Human Action Recognition." Revue d'Intelligence Artificielle 38, no. 1 (2024): 11-23.

[25] Al-muqarm, Abbas M. Ali, Yousif Mudhafar, Aboothar Mahmood Shakir, Murtada Kazem, Ruqaya Abdel-Yahiya, and Baqer Saad Abdul Zahra. "Low-Cost Smart Learning with Moodle-Based Raspberry Pi 4 for University Students." In 2023 6th International Conference on Engineering Technology and its Applications (IICETA), IEEE, 2023, 603-608.

[26] Abdulnabi, Saif Hasan, Yousif Samer Mudhafar, Ali Abdulhassan Kadhim, Mohammad Baqer Mahdi, and Hassan Hadi Sojar. "Neural Network-Based System Identification: A Comprehensive FPGA Design and Implementation." In 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), IEEE, 2024, 1-7.