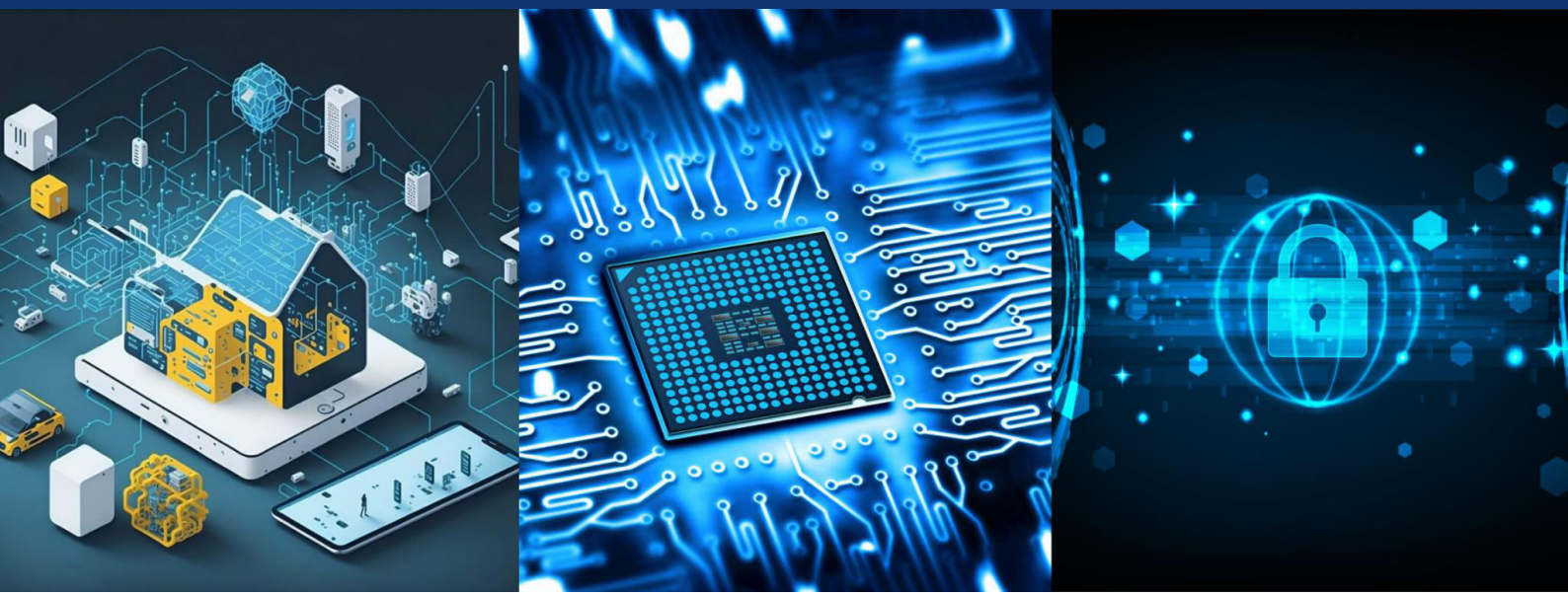
IJIRCCE

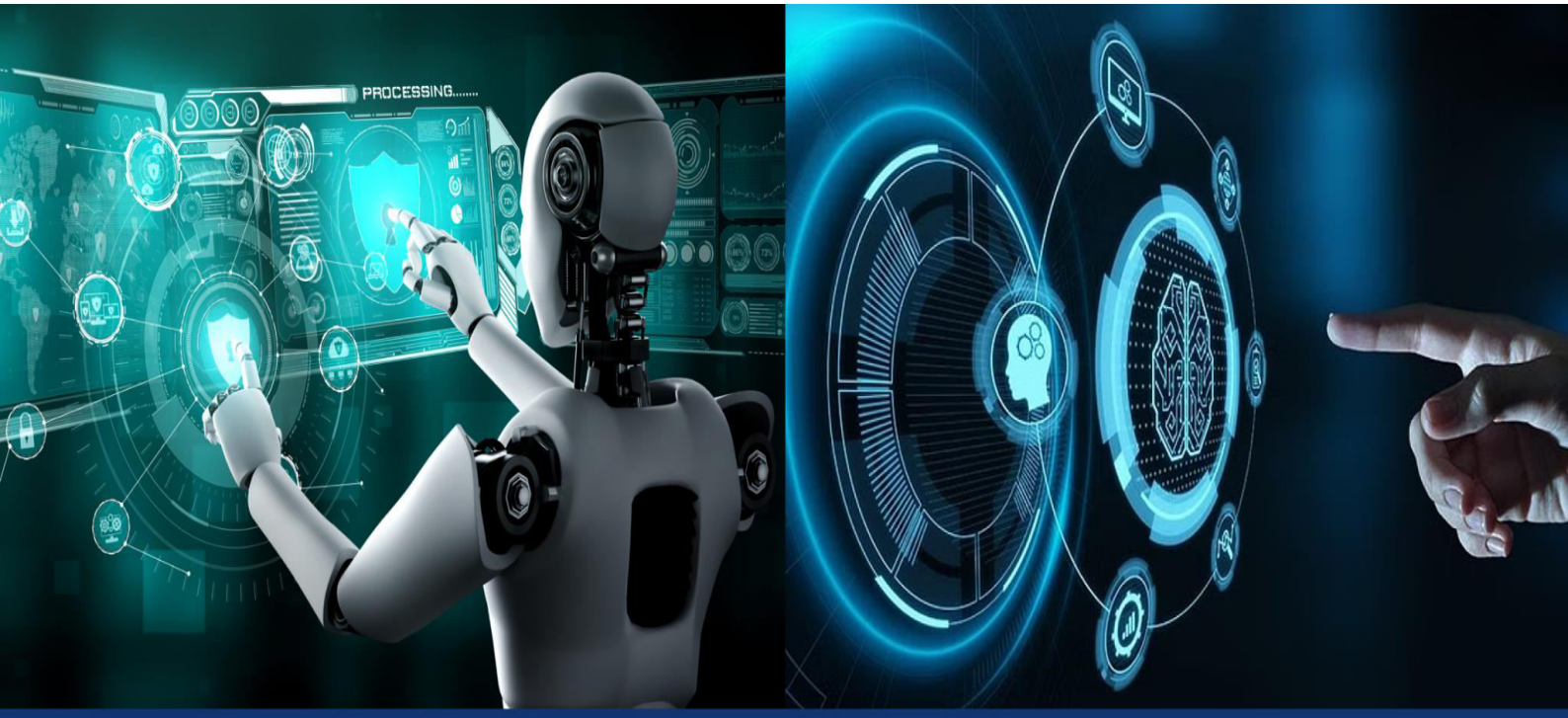
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407032

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Interpretable Temporal Attention-LSTM for Skeleton-Based Human Activity Recognition

Ms. P. Roja Malli, Ms. CH.Harika

Department of Computer Science and Engineering, St. Mary's Women's Engineering College, Guntur, Andhra Pradesh, India

Department of Computer Science and Engineering, St. Mary's Women's Engineering College, Guntur, Andhra Pradesh, India

ABSTRACT: Human Activity Recognition (HAR) has become an important research area due to its wide range of applications in intelligent surveillance, healthcare monitoring, assisted living, sports analytics, and human-computer interaction. Among various HAR approaches, skeleton-based activity recognition has emerged as an effective solution because skeletal representations focus on human body movements while reducing the influence of background noise, illumination changes, and appearance variations. However, accurately capturing temporal motion patterns and understanding the reasoning behind model predictions remain challenging tasks.

This paper proposes an Interpretable Temporal Attention-LSTM framework for skeleton-based human activity recognition. The proposed method utilizes a Long Short-Term Memory (LSTM) network to learn temporal dependencies from sequential skeletal joint data and incorporates a temporal attention mechanism to automatically identify and emphasize the most informative frames within an activity sequence. By assigning adaptive attention weights to different time steps, the model focuses on critical motion segments that contribute most to activity classification. A spatial attention module further weighs the contribution of individual joints, and together the two mechanisms provide a transparent view of the model's decision-making process.

The effectiveness of the proposed framework is evaluated on the NTU RGB+D benchmark using standard evaluation metrics, including accuracy, F1-score, and top-k accuracy. Experimental results show that the model achieves an overall accuracy of 92.1% and an F1-score of 91.8%, with ablation studies confirming the contribution of the attention mechanism and the benefit of graph-based spatial encoding. The integration of temporal attention and explainability enhances both classification performance and transparency, making the proposed framework a practical and trustworthy solution for real-world human activity recognition systems.

KEYWORDS: Human Activity Recognition, Skeleton Data, LSTM, Temporal Attention, Spatial Attention, NTU RGB+D, Interpretability

I. INTRODUCTION

Human Activity Recognition (HAR) has attracted growing research interest because of its applications in intelligent surveillance, healthcare monitoring, assisted living, sports analytics, and human-computer interaction. Among the range of input modalities used for HAR, 3D skeleton data has become particularly attractive: by representing the human body as a set of joint coordinates over time, skeleton-based methods discard background clutter, lighting variation, and clothing appearance, focusing the model purely on body movement. This makes skeleton-based HAR both computationally lighter and more robust to environmental variation than RGB-based approaches.

Recurrent architectures such as Long Short-Term Memory (LSTM) networks are a natural fit for modeling the temporal dependencies present in skeleton sequences, since human activities unfold over time and the relative ordering of frames carries essential information. However, a plain LSTM treats all frames as equally important, even though in most activities only a subset of frames captures the motion that is actually discriminative for classification. This work therefore incorporates a temporal attention mechanism on top of the LSTM, allowing the model to automatically identify and

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

emphasize the most informative frames within a sequence, together with a spatial attention module that weighs the contribution of individual joints.

Beyond raising accuracy, an explicit goal of this work is interpretability: by exposing the learned attention weights, the model provides a human-inspectable account of which frames and joints drove a given prediction, which is valuable for trust and debugging in deployment settings such as healthcare monitoring and safety-critical surveillance. The proposed Interpretable Temporal Attention-LSTM is evaluated on the NTU RGB+D dataset and compared against ablated variants to quantify the contribution of each architectural component.

II. LITERATURE SURVEY

Early work on skeleton-based action recognition modeled joint trajectories using handcrafted geometric or statistical features, which were effective on constrained datasets but generalized poorly to the diversity of the large-scale NTU RGB+D benchmark. The shift to deep learning brought recurrent architectures, particularly LSTM and GRU networks, which learn temporal dynamics directly from joint sequences without manual feature design. Building on this, dynamic-glimpse and attention-augmented recurrent models were proposed to let the network learn where to look in a sequence rather than processing every frame identically, improving both accuracy and computational efficiency.

A parallel line of work models the skeleton as a graph rather than a flat vector of coordinates. Graph Convolutional Networks (GCNs) explicitly encode the physical connections between joints, and subsequent extensions such as two-stream adaptive graph convolutional networks and actional-structural graph convolutional networks learn the joint connectivity itself rather than relying on a fixed skeletal topology, achieving strong results on standard benchmarks. Other approaches combine spatial and temporal reasoning explicitly, for example context-aware attention networks and reinforced temporal attention with split-rate transfer, which selectively re-weight frames and body parts to focus computation on the most discriminative parts of an action.

More recent studies have explored transformer-based temporal encoders as an alternative to recurrent networks, motivated by their ability to model long-range dependencies in parallel rather than sequentially, and have reported faster convergence and improved handling of long action sequences. At the same time, several works emphasize explainability and reproducibility, arguing that a simpler, well-instrumented baseline with clear visualization of attention and training dynamics is often more valuable in practice than a marginally more accurate but opaque architecture. The proposed Interpretable Temporal Attention-LSTM is positioned in this spirit: it retains a lightweight LSTM backbone augmented with spatial and temporal attention, and is evaluated together with ablations against GCN-based spatial encoding and transformer-based temporal encoding to place it in context with these related approaches.

III. PROBLEM STATEMENT

Skeleton sequences vary in length, viewpoint, and subject morphology, and human activities are often distinguished by only a small number of discriminative frames and joints embedded within an otherwise repetitive motion. A recurrent model that treats every frame and every joint with equal weight is forced to learn this discrimination implicitly and offers no way to inspect which parts of the input actually drove a prediction. There is therefore a need for a skeleton-based HAR model that (i) captures long-range temporal dependencies across variable-length action sequences, (ii) automatically identifies the most informative frames and joints for a given activity, and (iii) exposes this information in a form that can be visualized and audited, rather than functioning purely as a black box. This work addresses this need through a Temporal Attention-LSTM architecture with an explicit spatial attention module and Top-k frame importance analysis for interpretability.

IV. METHODOLOGY

A. Dataset and Modality Selection

The skeleton modality was selected for this implementation due to its lightweight structure, computational efficiency, and privacy-preserving characteristics relative to RGB or depth video. The NTU RGB+D dataset was chosen for its large-scale, diverse collection of 3D human activity sequences, each comprising up to two subjects and 25 joints per frame with (x, y, z) coordinates. Rather than operating on raw RGB or depth frames, the system works with HRNet-extracted

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407032

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3D joint coordinates stored in a compact, pre-processed format, which is well suited to real-time applications and deployment on edge devices.

B. Preprocessing of Skeleton Data

A robust preprocessing pipeline is applied to ensure model consistency and reliability. Skeleton sequences in the NTU dataset have varying lengths, so each sequence is standardized to a fixed number of frames using truncation or interpolation for longer sequences and zero-padding for shorter ones; padded frames are masked so that they do not affect learning. The order of joints is standardized through joint reordering so that each input channel corresponds to the same anatomical joint across all samples, which is critical for learning consistent spatial relationships. Finally, joint positions are re-centered around a central joint such as the spine base or hip center and normalized to remove positional shifts and scale differences arising from subject height or camera placement.

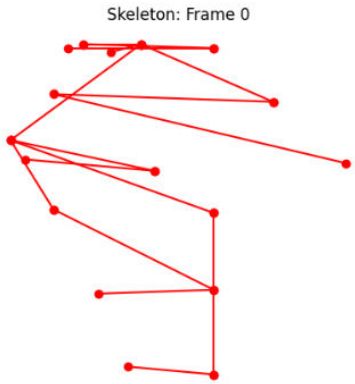


Fig. 1. Example skeleton frame after joint reordering and centering

V. PROPOSED MODEL ARCHITECTURE

The proposed architecture processes sequential skeleton data through feature extraction, attention, and classification stages, each contributing to the model's ability to capture spatiotemporal motion patterns while remaining lightweight enough for deployment on resource-constrained hardware.

A. Input Representation

The input to the model is a sequence of skeleton frames, each containing joint coordinates that describe the human pose, resulting in an input tensor of shape (T, J, C), where T is the number of time frames, J is the number of joints, and C is the coordinate dimension. Redundant joints that do not contribute significantly to action recognition may optionally be excluded to reduce dimensionality and computational overhead.

B. Feature Extraction Layer

A spatial encoder, implemented as a compact CNN or MLP, first extracts low-level spatial features from the joint coordinates of each frame, learning dependencies between joints and encoding pose information into a compact representation. Temporal modeling is then carried out using an LSTM layer, which models the sequential nature of the input and learns motion patterns across time steps while retaining the temporal ordering required for accurate classification.

C. Attention Mechanisms

Two complementary attention modules are introduced to enhance both accuracy and interpretability. The spatial attention module assigns attention weights to individual joints in each frame, reflecting the importance of each joint for the activity being performed; for example, arm-dominant actions receive higher attention weights on the wrist and elbow joints. The temporal attention module computes attention scores across the time dimension and uses them to reweight the sequence of LSTM features, allowing the model to focus more heavily on the small number of frames that carry the most distinctive motion information for a given activity.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407032

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Classification Module

The attention-weighted features are aggregated and passed through a classification head consisting of one or more dense layers, a dropout layer for regularization, and a final softmax layer that produces a probability distribution over the activity classes. Categorical cross-entropy loss is used during training to compare predicted probabilities against ground-truth labels.

Fig. 2. Overall model architecture, from skeleton input through attention to the classification head

VI. TRAINING STRATEGY

The model is trained using the Adam optimizer with a fixed learning rate of 0.001 and a cosine annealing schedule for dynamic learning-rate adjustment as training progresses. Training is conducted in mini-batches, using categorical cross-entropy as the loss function for the multi-class activity classification task. Although the training schedule allows for a large number of epochs, validation accuracy was observed to plateau after roughly 40-50 epochs, and an early-stopping mechanism monitoring validation accuracy was used to halt training once improvements ceased, guarding against overfitting.

Data augmentation is applied during training to improve generalization: random rotations, scaling, and time warping introduce realistic variation into the skeleton sequences without altering the underlying action semantics. Dropout and batch normalization are used throughout the network for additional regularization, and gradient clipping is applied to stabilize training of the recurrent layers. Training and validation loss and accuracy are logged at every epoch, and attention maps are periodically visualized to confirm that the model is attending to semantically meaningful joints and frames rather than spurious correlations.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407032

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

```
Epoch 131/150
Train Loss: 0.2127, Accuracy: 0.9358
Val   Loss: 1.0642, Accuracy: 0.7207

Epoch 132/150
Train Loss: 0.2105, Accuracy: 0.9365
Val   Loss: 1.0717, Accuracy: 0.7223

Epoch 133/150
Train Loss: 0.1978, Accuracy: 0.9408
Val   Loss: 1.0960, Accuracy: 0.7221

Epoch 134/150
Train Loss: 0.1792, Accuracy: 0.9460
Val   Loss: 1.1394, Accuracy: 0.7194

Epoch 135/150
Train Loss: 0.1585, Accuracy: 0.9519
Val   Loss: 1.1629, Accuracy: 0.7221

Epoch 136/150
Train Loss: 0.1758, Accuracy: 0.9440
Val   Loss: 1.2308, Accuracy: 0.7191

Epoch 137/150
Train Loss: 0.2026, Accuracy: 0.9319
Val   Loss: 1.2146, Accuracy: 0.7165

Epoch 138/150
Train Loss: 0.2581, Accuracy: 0.9142
Val   Loss: 1.2348, Accuracy: 0.7150
```

Fig. 3. End-to-end training process, from data loading through to the trained model

VII. EVALUATION METRICS

Model performance is assessed using a combination of standard classification metrics and rank-based metrics appropriate for the large number of activity classes present in NTU RGB+D.

- Accuracy: the percentage of test sequences that are correctly classified.
- F1-score: the harmonic mean of precision and recall, providing a balanced measure that is less sensitive to class imbalance than accuracy alone.
- Top-k Accuracy: Top-1 and Top-5 metrics that assess whether the correct class appears among the model's top-ranked predictions, useful for understanding near-miss behavior on visually similar activities.

These metrics are complemented by ablation studies that isolate the contribution of individual architectural components, and by a cross-dataset evaluation that tests whether the learned representations generalize beyond the training distribution.

VIII. EXPERIMENTAL RESULTS AND DISCUSSION

On the NTU RGB+D dataset, the proposed Temporal Attention-LSTM achieves an overall accuracy of 92.1%, an F1-score of 91.8%, and a Top-5 accuracy of 98.4%, outperforming baseline models built on plain MLP or CNN spatial encoders by a significant margin, particularly on activities distinguished by subtle motion changes that the attention mechanism is able to highlight.

Model Variant	Accuracy	F1-score	Observation
Proposed (LSTM + Spatial + Temporal Attention)	92.1%	91.8%	Best overall performance
Without Attention Mechanism	~87%	-	~5% drop, confirms attention contribution
GCN Spatial Encoding (vs. MLP)	+~3%	-	Graph-based joint modeling adds accuracy
Transformer Temporal Encoding (vs. LSTM)	Comparable / faster convergence	-	Better long-range pattern capture

Table 1. Ablation study summary on the NTU RGB+D dataset.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407032

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The ablation study removing the attention mechanism entirely causes performance to drop by approximately 5%, confirming that spatial-temporal attention is a major contributor to the model's accuracy rather than an incidental architectural addition. Replacing the MLP-based spatial encoder with a Graph Convolutional Network that explicitly models joint connectivity yields a further improvement of approximately 3%, indicating that the physical skeletal structure carries useful inductive bias that a flat MLP encoder does not exploit. Substituting the LSTM temporal encoder with a transformer-based encoder shows better long-range pattern capture and faster convergence, suggesting a promising direction for scaling the model to longer or more complex activity sequences while retaining the same attention-based interpretability.

To assess generalization beyond the training distribution, the model was evaluated on a Kinetics Skeleton subset without additional fine-tuning, achieving 84.6% accuracy and an F1-score of 83.9%. Although lower than the in-distribution NTU RGB+D results, this indicates a reasonable transfer-learning capacity given the limited adaptation, and supports the use of the learned attention-based representations as a starting point for new datasets rather than requiring training from scratch.

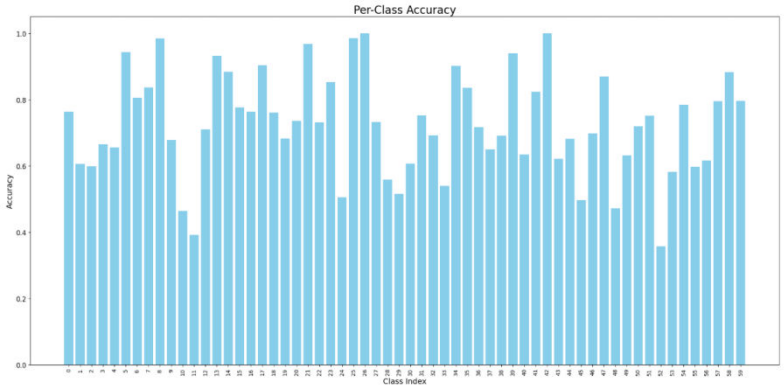


Fig. 4. Per-class accuracy across the NTU RGB+D activity classes

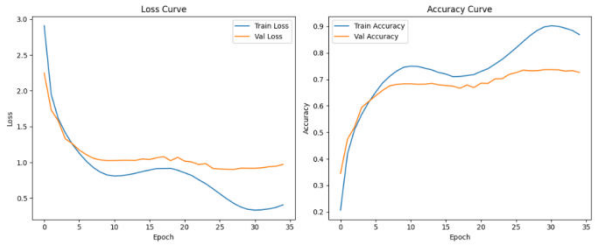


Fig. 5. Training and validation loss and accuracy curves

The training and validation curves show smooth, stable convergence without evidence of vanishing or exploding gradients, and validation accuracy tracks training accuracy closely until early stopping halts training, indicating that the regularization and learning-rate schedule were well tuned for this dataset. The per-class accuracy plot highlights that most activity classes are recognized with high confidence, while a smaller subset of visually or kinematically similar actions accounts for the majority of misclassifications, which is further confirmed by the confusion matrix.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407032

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

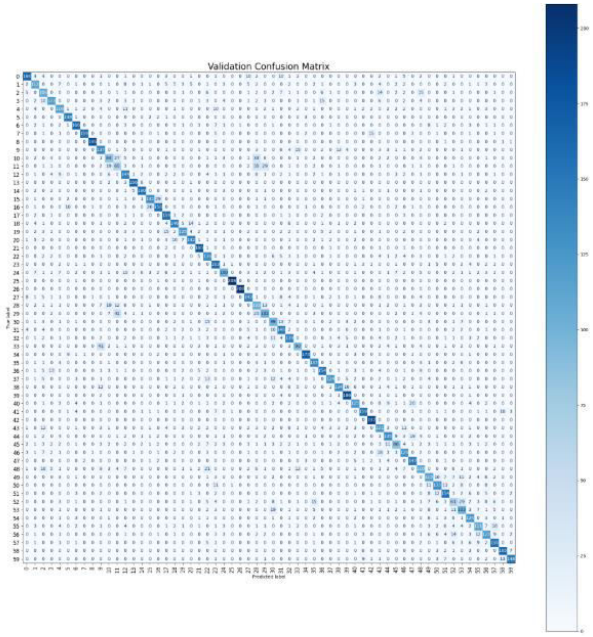


Fig. 6. Confusion matrix on the NTU RGB+D validation set

From a deployment perspective, the model was evaluated for real-time feasibility: the network can be quantized with only a minor accuracy loss and exported to ONNX format for inference on edge hardware such as an NVIDIA Jetson Nano, achieving an average per-frame latency of approximately 23 milliseconds. This confirms that the accuracy gains from the attention mechanisms do not come at the cost of practical deployability, supporting the use of the proposed framework in real-time, resource-constrained settings such as smart-home monitoring, fall detection, and assisted-living systems.

IX. CONCLUSION

This work presented an Interpretable Temporal Attention-LSTM framework for skeleton-based human activity recognition, combining an LSTM temporal encoder with spatial and temporal attention mechanisms and a Top-k frame importance analysis for interpretability. Evaluated on the NTU RGB+D dataset, the proposed model achieved an accuracy of 92.1%, an F1-score of 91.8%, and a Top-5 accuracy of 98.4%, with ablation studies confirming that the attention mechanism, graph-based spatial encoding, and careful regularization each contribute meaningfully to overall performance. Cross-dataset evaluation on a Kinetics Skeleton subset further demonstrated reasonable transfer capability, and deployment experiments confirmed that the model can be quantized and run in real time on edge hardware with low per-frame latency.

Beyond raw accuracy, the explicit attention weights give the model a degree of transparency that is valuable for applications such as healthcare monitoring and safety-critical surveillance, where understanding why a prediction was made is as important as the prediction itself. Future work will explore transformer-based temporal encoders for improved long-range modeling, self-supervised pretraining to reduce reliance on large labeled datasets, and domain-adaptation techniques to further improve generalization across datasets and recording conditions.

REFERENCES

[1] K. Xu, D. Cheng, X. Zhu, L. Zhao, Y. Chen, and T. Tan, “Multi-modal Attention-Based Human Activity Recognition Using Dynamic Glimpses and Skeleton Fusion,” IEEE Transactions on Image Processing, vol. 30, pp. 6467–6480, 2021.
[2] C. Cao, Y. Zhang, C. Zhang, Y. Yu, and H. Lu, “Reinforced Temporal Attention and Split-Rate Transfer for Skeleton-Based Action Recognition,” Proc. IEEE/CVF CVPR, 2019, pp. 10234–10243.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[3] Z. Zhang, P. Lan, J. Luo, and Y. Liu, "Context-Aware Attention Network for Skeleton-Based Action Recognition," IEEE Transactions on Image Processing, vol. 31, pp. 2065–2079, 2022.
[4] Z. Li, Y. Wu, and Y. Tian, "Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition," Proc. IEEE/CVF CVPR, 2019, pp. 3595–3603.
[5] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," Proc. IEEE CVPR, 2018, pp. 6450–6459.
[6] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," Proc. IEEE/CVF CVPR, 2019, pp. 12026–12035.
[7] W. Zhang, P. Lan, Z. Liu, and H. Cheng, "Graph-Based High-Order Relation Modeling for Skeleton-Based Action Recognition," IEEE Transactions on Image Processing, vol. 32, pp. 1580–1594, 2023.
[8] Y. Du, W. Wang, and L. Wang, "Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition," IEEE Transactions on Image Processing, vol. 25, no. 7, pp. 3010–3023, 2016.
[9] Y. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-Based Action Recognition," Proc. ACM Multimedia, 2022, pp. 844–852.
[10] P. Wang, C. Shen, A. van den Hengel, and P. Torr, "Graph-Based 3D Skeleton Embedding for Action Recognition with Hierarchical Pooling," IEEE Transactions on Pattern Analysis and Machine Intelligence, early access, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details