Interpretable Human Action Recognition: A CNN-GRU Approach with Gradient-weighted Class Activation Mapping Insights

Md. Sabir Hossain [1] **Mufti Mahmud** [*1 2] **Md. Mahfuzur Rahman** [*1 3]

Abstract

Human Action Recognition (HAR) is essential in applications like healthcare, surveillance, and smart environments, where reliable and interpretable decision-making is critical. While Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs) effectively model spatial and temporal patterns, their black-box nature limits transparency in safety-sensitive domains. This study introduces an interpretable HAR framework combining a CNN-GRU architecture with Gradient-weighted Class Activation Mapping (Grad-CAM). The CNN captures frame-wise spatial features, GRUs model temporal dynamics, and a 3D convolution bridges spatial-temporal abstraction. Grad-CAM provides frame-level heatmaps to visualize model rationale. Evaluated on 10 diverse classes from the UCF101 dataset, our model achieved 96.50% accuracy and outperformed several standard deep models across precision, recall, and F1 metrics. Visual analysis of correct and incorrect cases confirms both model reliability and interpretability. The framework offers a robust and transparent solution for real-time HAR in critical domains.

1. Introduction

Human Activity Recognition (HAR) has gained momentum across diverse domains such as healthcare, surveillance, and smart environments due to its potential to enable automatic and accurate behavior classification from video or sensor data (Abdellatef et al., 2025). Traditional approaches rely heavily on handcrafted features and classical machine learning algorithms, which often struggle with generalization to complex, real-world settings (Oleh et al., 2024).

Recent advances in deep learning have allowed HAR systems to learn spatial and temporal patterns directly from raw inputs. CNNS effectively capture local spatial representations, while recurrent units such as GRUs and LSTMs model temporal dependencies in video streams (Cheng & Xu, 2024). Hybrid models, particularly CNN-GRU architectures, have demonstrated superior performance on benchmark datasets by leveraging the strengths of both modalities (Verma et al., 2022; Cheng & Xu, 2024).

However, the black-box nature of these models remains a barrier to deployment in sensitive applications. Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM have emerged as promising solutions to visualize and interpret the internal reasoning of neural networks (Mahmoudi et al., 2023; Aquino et al., 2023). These methods enhance transparency and support validation by highlighting spatiotemporal cues contributing to predictions.

This work presents an interpretable HAR framework combining a CNN-GRU model with Grad-CAM-based post-hoc interpretability. Evaluated on a selected subset of the UCF101 dataset, our model achieves competitive classification performance while providing frame-level visual explanations that uncover meaningful activity cues. The results demonstrate the feasibility of integrating accurate recognition with actionable interpretability, advancing HAR toward real-world, trust-sensitive applications.

2. Literature Review

Deep learning has advanced HAR through end-to-end learning, but its black-box nature raises trust issues in sensitive areas like healthcare. This drives interest in XAI to improve interpretability and reliability in real-world applications.

Deep learning has become the foundation of modern HAR, allowing automatic feature extraction without hand-crafted

[1]Information and Computer Science Department, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia [2]SDAIA-KFUPM Joint Research Center for AI and Interdisciplinary Research Center for Bio Systems and Machines, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia [3]Interdisciplinary Research Center for Intelligent Secure Systems, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia. Correspondence to: Mufti Mahmud <muftimahmud@gmail.com, mufti.mahmud@kfupm.edu.sa>, Md Mahfuzur Rahman <mdmahfuzur.rahman@kfupm.edu.sa>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. 4^{th} MusIML Workshop, 2025.

engineering. CNNs are widely used for spatial-temporal representation learning, with hybrid CNN-RNN models (e.g., CNN-GRU or CNN-LSTM) offering superior performance in modeling sequential dependencies (Chandramouli et al., 2024; Cheng & Xu, 2024). For example, hierarchical CNN architectures have shown high accuracy across datasets like UCI-HAR and KU-HAR, while hybrid structures outperform standalone models in recognizing complex motion patterns (Verma et al., 2022; Abdellatef et al., 2025). Recent trends include leveraging self-supervised and transfer learning to address data scarcity in HAR, where pretrained models on large-scale unlabeled corpora have improved generalization on downstream tasks. Multi-modal fusion and context-aware designs have further boosted robustness in unconstrained environments (Oleh et al., 2024). In our work, we adopt a CNN-GRU structure optimized for video-based action recognition, refining it by removing post-convolutional dropout to improve training stability without compromising spatial-temporal modeling (Taghiyev, 2021).

With increasing demand for transparency, XAI methods have been explored to interpret deep HAR models. Techniques like SHAP and LRP offer post-hoc explanations but face challenges with high-dimensional and temporally structured data. SHAP can misattribute importance in sequential contexts, while LRP often struggles with deep recurrent or 3D convolutional models (Van Zyl et al., 2024; Sun et al., 2022). Attention-based approaches provide intuitive heatmaps but may lack contextual granularity, failing to capture cross-modal or long-range dependencies (Li et al., 2018; Hao et al., 2022; Moniruzzaman et al., 2021). Grad-CAM has emerged as a practical solution for visualizing spatial and temporal model attention in video sequences. It highlights influential regions in frames, offering clearer interpretation of decision logic in HAR tasks (Jayamohan & Yuvaraj, 2025; Alam et al., 2024). By integrating Grad-CAM into our CNN-GRU framework, we aim to bridge high performance with interpretability, enabling actionable insights into model reliability and misclassification behavior.

3. Methodology

We propose an interpretable human action recognition pipeline that integrates a hybrid CNN-GRU architecture with Grad-CAM-based visual interpretability. As illustrated in Figure 1, the framework processes uniformly sampled 10-frame video clips from the UCF101 dataset, performs classification via spatio-temporal deep learning, and generates frame-wise heatmaps to highlight decision-critical regions. The goal is to couple accurate classification with actionable explanations, particularly useful in trust-sensitive applications.

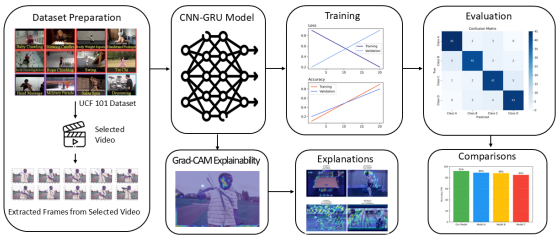

Figure 1. Overview of the proposed pipeline combining CNN-GRU modeling with Grad-CAM explanations.

3.1. CNN-GRU-based Architecture

The CNN-GRU architecture (Figure 2) processes 10-frame video sequences, where each frame (128×192 RGB) is passed through a shared CNN encoder to extract spatial features. This encoder consists of a block of three Conv2D layers interleaved with Batch Normalization, MaxPooling, and Dropout, applied independently to each frame. The extracted features are then stacked along the temporal axis and passed through a Conv3D layer with 50 filters to capture local spatio-temporal dynamics.

The resulting tensor is reshaped and fed into a two-stage GRU block with 32 and 50 units, which models long-range temporal dependencies across the sequence. The final layers consist of a Dense layer with 100 neurons (ReLU activation) followed by a Softmax layer that outputs probabilities over 10 action classes. This compact hybrid design balances spatial, short-term, and long-term temporal modeling while maintaining computational efficiency. The overall architectural flow is illustrated in Figure 2.

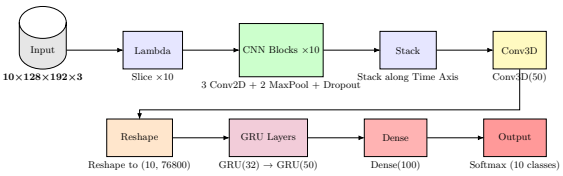

Figure 2. CNN-GRU model architecture for spatio-temporal action recognition.

3.2. Grad-CAM-based Explainability

To visualize model attention, we apply Grad-CAM, which highlights the spatial regions contributing most to the predicted action. As illustrated in Figure 3, Grad-CAM computes the class-specific importance weights α_k^c by averaging the gradients of the score y^c over feature maps A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

These weights are linearly combined and passed through a ReLU to form the heatmap:

$$L^c = \text{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k\right) \tag{2}$$

The resulting heatmaps are upsampled and superimposed on the input frames, producing a sequence of interpretable overlays. These Grad-CAM visualizations are computed frame-by-frame and compiled into videos, allowing users to inspect how attention evolves over time.

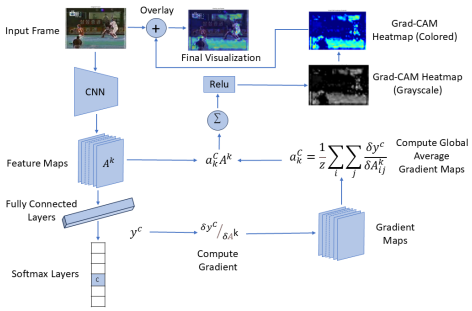

Figure 3. Grad-CAM pipeline showing class-discriminative regions across frames.

When applied to correctly predicted actions such as *BaseballPitch*, Grad-CAM consistently emphasized motion-relevant body regions, such as arm swing and body orientation. This reinforces the alignment between model focus and semantic cues, increasing user trust and model transparency.

4. Experimental Results and Discussion

4.1. Dataset and Experimental Setup

We evaluated our model on a 10-class subset of the UCF101 dataset (Soomro et al., 2012), a standard benchmark for video-based action recognition. Ten classes were randomly selected to ensure motion diversity, with 100 videos per class (1,000 total). From each video, 10 uniformly spaced RGB frames were extracted and resized to 128×192 pixels.

The dataset was split 80:20 for training and testing, with one-hot encoded labels. Training was conducted using TensorFlow/Keras on Kaggle with dual NVIDIA Tesla T4 GPUs, and results were validated locally on an Intel i7 CPU with 32GB RAM.

4.2. Performance Analysis

Our CNN-GRU model was trained over 20 epochs with smooth convergence observed in training and validation loss/accuracy curves (Fig. 4). On the test set, it achieved

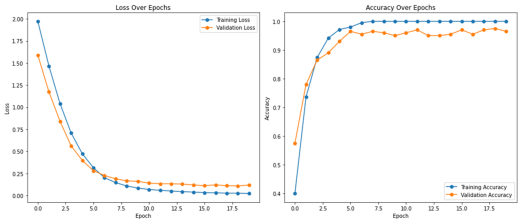

Figure 4. Training and validation loss and accuracy.

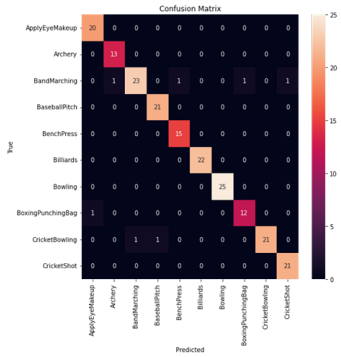

Figure 5. Confusion matrix for 10-class HAR classification on UCF101 subset.

96.5% accuracy, demonstrating robust generalization. Figure 5 presents the confusion matrix, showing strong class-wise prediction accuracy. Minor confusion occurred in visually similar actions, but overall classification was consistent and reliable.

4.3. Grad-CAM Visualization and Insights

Figure 6 shows Grad-CAM visualizations across a variety of action classes, illustrating the model's attention to semantically relevant regions during prediction. These heatmaps help interpret how the model arrives at its decisions.

For correctly classified examples such as *Billiards*, Grad-CAM highlights key objects like the cue and balls (Fig. 7), confirming that the model focuses on meaningful visual cues. In contrast, in a misclassified instance where *Bowling* was predicted as *CricketBowling*, the attention maps are dispersed and misaligned (Fig. 8), suggesting difficulties in distinguishing visually similar actions. These frame-level visual explanations demonstrate the interpretability and diagnostic value of Grad-CAM, offering insights into both the model's strengths and its limitations.

4.4. Comparison with Deep Learning Baselines

We compared our model with several well-established CNN-based baselines, including ResNet50, InceptionV3, MobileNet, XceptionNet, VGG16, VGG19 and DenseNet121, all evaluated on the same 10-class subset of the UCF101

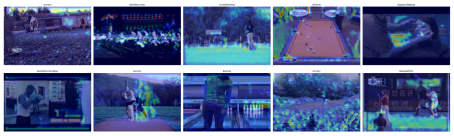

Figure 6. Grad-CAM heatmaps highlighting key spatial regions influencing action recognition decisions.

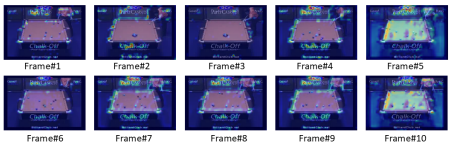

Figure 7. Grad-CAM heatmaps for correctly classified *Billiards*.

dataset using consistent training settings. As shown in Table 1, CNN-GRU outperformed all baselines with 96.50% accuracy and balanced precision/recall. While XceptionNet and DenseNet121 reached 94.00% and 92.00% respectively, static CNNs like ResNet50 and VGG variants underperformed due to their lack of temporal modeling.

4.5. Comparative Analysis with SOTA Benchmarks

To further validate our model's effectiveness, we compared it with top-performing models from the UCF101 leaderboard on PapersWithCode[1], ensuring a fair comparison by evaluating all models, including ours, using only the original UCF101 data without external augmentation. These models include A2-Net (96.40%), I3D-LSTM (95.10%), TS-LSTM (94.10%), Two-stream+LSTM (88.60%), and HalluciNet (79.83%). As seen in Table 2, our model achieves superior accuracy (96.50%) without requiring complex multi-stream inputs or hallucinated representations.

The consistent outperformance of CNN-GRU reflects the advantages of fusing spatial and temporal modeling, further supported by interpretability through Grad-CAM. This makes our approach not only accurate but also transparent and deployable in real-world scenarios such as surveillance, healthcare monitoring, and assistive systems.

[1] https://paperswithcode.com/sota/action-recognition-in-videos-on-ucf101?tag_filter=8%2C3

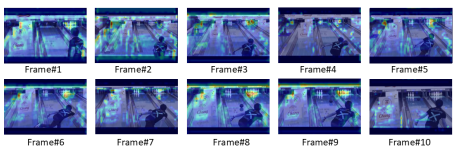

Figure 8. Grad-CAM heatmaps of misclassified *Bowling*.

Table 1. Comparison of classification performance across models.

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
CNN-GRU	0.9650	0.9657	0.9650	0.9644
RESNET50	0.3250	0.2998	0.3250	0.2397
INCEPTIONV3	0.8950	0.9052	0.8950	0.8939
MOBILENET	0.8800	0.9070	0.8800	0.8827
XCEPTIONNET	0.9400	0.9498	0.9400	0.9402
VGG16	0.0650	0.0051	0.0650	0.0094
VGG19	0.0650	0.0042	0.0650	0.0079
DENSENET121	0.9200	0.9459	0.9200	0.9253

Table 2. Benchmarking against top UCF101 models (single-stream).

MODEL	ACCURACY (%)
CNN-GRU [THIS WORK]	96.50
A2-NET (RESNET-50) (CHEN ET AL., 2018)	96.40
I3D-LSTM (WANG ET AL., 2019)	95.10
TS-LSTM (MA ET AL., 2019)	94.10
TWO-STREAM+LSTM (YUE-HEI NG ET AL., 2015)	88.60
HALLUCINET (RESNET-50) (PARMAR & MORRIS, 2021)	79.83

5. Conclusion and Future Directions

This work introduced an interpretable deep learning framework for HAR that combines a CNN-GRU hybrid architecture with Grad-CAM. Evaluated on a 10-class subset of the UCF101 dataset, the model achieved 96.50% accuracy, demonstrating its ability to capture complex spatiotemporal patterns in video data. The use of Grad-CAM provided class-discriminative heatmaps that visually confirmed the model's attention to semantically meaningful regions, offering transparency in both correct and incorrect predictions. This interpretability is especially valuable in high-stakes applications where model trust is critical, such as healthcare, surveillance, and human–computer interaction. Moreover, the model's strong performance using limited annotated data highlights its suitability for resource-constrained environments.

Future work should enhance interpretability beyond Grad-CAM using finer-grained or multimodal XAI to explain spatial and temporal features. Combining video with inertial or depth data can boost real-world robustness. Attention mechanisms and Transformers may improve long-range temporal modeling. Cross-domain generalization and domain adaptation are essential for robust, personalized HAR deployment.

Acknowledgements

This research was fully supported by the Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS) through the Deanship of Research, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia (grant no. #INSS2516).

References

Abdellatef, E., Al-Makhlasawy, R. M., and Shalaby, W. A. Detection of human activities using multi-layer convolutional neural network. *Scientific Reports*, 15(1):7004, 2025.

Alam, M. T., Acquaah, Y. T., and Roy, K. Image-based human action recognition with transfer learning using grad-cam for visualization. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 117–130. Springer, 2024.

Aquino, G., Costa, M. G. F., and Filho, C. F. F. C. Explaining and visualizing embeddings of one-dimensional convolutional models in human activity recognition tasks. *Sensors*, 23(9):4409, 2023.

Chandramouli, N. A., Natarajan, S., Alharbi, A. H., Kannan, S., Khafaga, D. S., Raju, S. K., Eid, M. M., and El-Kenawy, E.-S. M. Enhanced human activity recognition in medical emergencies using a hybrid deep cnn and bi-directional lstm model with wearable sensors. *Scientific Reports*, 14(1):30979, 2024.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., and Feng, J. Aˆ 2-nets: Double attention networks. *Advances in neural information processing systems*, 31, 2018.

Cheng, C. and Xu, H. A 3d motion image recognition model based on 3d cnn-gru model and attention mechanism. *Image and Vision Computing*, 146:104991, 2024.

Hao, Y., Wang, S., Cao, P., Gao, X., Xu, T., Wu, J., and He, X. Attention in attention: Modeling context correlation for efficient video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7120–7132, 2022.

Jayamohan, M. and Yuvaraj, S. A novel human action recognition using grad-cam visualization with gated recurrent units. *Neural Computing and Applications*, pp. 1–16, 2025.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, D., Yao, T., Duan, L.-Y., Mei, T., and Rui, Y. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 21(2):416–428, 2018.

Ma, C.-Y., Chen, M.-H., Kira, Z., and AlRegib, G. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, 71:76–87, 2019.

Mahmoudi, S. A., Amel, O., Stassin, S., Liagre, M., Benkedadra, M., and Mancas, M. A review and comparative study of explainable deep learning models applied on action recognition in real time. *Electronics*, 12(9):2027, 2023.

Moniruzzaman, M., Yin, Z., He, Z., Qin, R., and Leu, M. C. Human action recognition by discriminative feature pooling and video segment attention model. *IEEE Transactions on Multimedia*, 24:689–701, 2021.

Oleh, U., Obermaisser, R., and Ahammed, A. S. A review of recent techniques for human activity recognition: Multimodality, reinforcement learning, and language models. *Algorithms*, 17(10):434, 2024.

Parmar, P. and Morris, B. Hallucinet-ing spatiotemporal representations using a 2d-cnn. *Signals*, 2(3):604–618, 2021.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sun, J., Lapuschkin, S., Samek, W., and Binder, A. Explain and improve: Lrp-inference fine-tuning for image captioning models. *Information Fusion*, 77:233–246, 2022.

Taghiyev, F. Video Action Recognition. https://www.kaggle.com/code/faridtaghiyev/video-action-recognition-ucf101, 2021. [Accessed 02-02-2025].

Van Zyl, C., Ye, X., and Naidoo, R. Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of grad-cam and shap. *Applied Energy*, 353:122079, 2024.

Verma, U., Tyagi, P., and Kaur, M. Single input single head cnn-gru-lstm architecture for recognition of human activities. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 10(2):410–420, 2022.

Wang, X., Miao, Z., Zhang, R., and Hao, S. I3d-lstm: A new model for human action recognition. In *IOP conference series: materials science and engineering*, volume 569-3, pp. 032035. IOP Publishing, 2019.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4694–4702, 2015.