International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

# Anomaly Detection in Surveillance Videos with Variable Temporal Observation

**Tanvi Prabhu[1], Shambhavi Rai[2], Sanya Sahu[3], Shubha Nandini B M[4], Dr. Sheshappa S N[5]**

[1,2,3,4] Department of Information Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, India

[5]Professor, Department of Information Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, India

tanvi211013@gmail.com

---

**Abstract -** Automated surveillance systems require not only post-event anomaly detection but also early action anticipation to enable proactive crime prevention and rapid emergency response. This project presents an end-to-end interpretable deep learning framework for early anomaly detection under variable temporal observations. The proposed architecture utilizes a dual-stream pipeline that decomposes surveillance video into spatial RGB frames and motion dynamics and uses Inception V3 to for feature extraction. To capture dynamic temporal dependencies, sliding sub-windows are processed using a Masked Bidirectional Gated Recurrent Unit (BiGRU) combined with a custom Temporal Attention Mechanism, which adaptively prioritizes crucial early predictive frames while filtering out uninformative or padded noise. Evaluated via 5-fold stratified cross-validation on a benchmark mini-UCF Crime dataset, the model achieves a strong early prediction accuracy of 73.6% at just 10% video observation, steadily increasing to 78.0% at 50% observation and reaching a peak of 79.6% at 75% observation. These results demonstrate the system's ability to reliably anticipate unfolding anomalies long before an event fully completes, offering a scalable solution for real-time intelligent surveillance.

***Key Words***: *early anomaly detection, surveillance videos, temporal attention, Bi-GRU, Inception V3*

## 1. INTRODUCTION

Nowadays, surveillance cameras are installed in many public places such as roads, shopping malls, colleges, railway stations and other crowded areas. These cameras generate a large amount of video data every day. Monitoring all these videos manually is difficult and time-consuming, especially when an incident needs to be identified quickly. Deep learning has made it possible to analyse video sequences automatically and recognize different human activities. Human Action Recognition (HAR) is useful in applications such as public safety monitoring, violence detection and identifying unusual activities in surveillance footage.

Traditional surveillance systems mainly depend on human observation or detect an event after it has already started. This can delay the response to potentially dangerous situations. Although deep learning models have achieved good results in human action recognition, recognizing an action from only a small portion of a video remains a challenge. In many situations, the complete video sequence may not be available when an early prediction is required. Therefore, there is a need for a system that can learn from the initial frames of a video and predict the possible action before the complete sequence is observed.

The main objectives of the project are:

1. To develop a deep learning-based system for anticipating human actions from partial surveillance video sequences.

2. To build a network with temporal attention to learn sequential dependencies and focus on informative frames.

3. To evaluate the model's performance using only 10%, 20%, 30%, 40%, 50% and 75% of the video.

## 2. LITERATURE SURVEY

3D Deep Learning (DL) has enabled machines to understand video as input. Due to constant surveillance, there is a massive amount of video data that cannot be processed manually, this is where human action recognition using DL comes into the picture. It helps in the monitoring of elderly patients for early fall detection, onset of violent behavior in crowded environments and improved human robot interaction. When we use 3D Deep Learning to recognize human activity in video sequences, the neural network must account for two aspects: spatial and temporal. It is well recognized that Convolution Neural Networks (CNN) are particularly effective in extracting spatial features. For temporal features, we have modern Recurrent Neural Networks (RNN) such as Long-Short Term Memory (LSTMs) and Gated Recurrent Units (GRUs). The more recent development of Transformers, specifically ViT (Vision Transformers), is competing with CNNs as it can handle both spatial and temporal features well, but with a higher computational cost.

Deep neural networks have a complex architecture which makes it hard to understand how it arrived at a specific output. This is known as the "black-box" nature of these models that makes interpretability difficult. Saliency maps, Grad-CAM (Gradient-weighted Class Activation Mapping) and SHAP (SHapley Additive exPlanations) are post-hoc explanations i.e. they explain the behavior of the model after it makes a prediction, whereas attention mechanisms provide intrinsic interpretability while the model is running.

The survey is divided into three subtopics: 3D Deep Learning models, Interpretability and xAI, and Attention Mechanisms.

### A. 3D Deep Learning Models

A major problem for HAR is cluttered backgrounds, different viewpoints, low resolution or partial occlusion in input video sequences. One way is to improve the data preprocessing; Rahayu et al.

uses a pure CNN based model with ReLU as an activation function to determine the optimal window size for HAR [1]. They determined that the optimal size was 16 with an accuracy of 94.08%. On the other hand, most researchers have shown that hybrid models outperform their standalone counterparts. Hussain et al. proposed one such model, which uses a pretrained Vit- Base-16 along with a multi-layer LSTM which achieves an accuracy of 73.714% on HMDB51 and 96.144% on UCF50 [2]. Another prominent way for HAR is skeleton-based action recognition, which employs Graph Convolutional Networks (GCN). Wang et al. proposes a model that treats two people in a video as a Merge and Split Graph (MSG) with 50 nodes rather than two independent graphs with 25 nodes [5]. Gao et al. introduced a novel Pose Correction and Channel Topology Refinement – GCN (PCCTR-GCN) that focuses on refinement during feature extraction [11]. They employ two modules:pose correction, which corrects posture coordinates in the input to reduce error, and channel topology refinement graph convolution, which dynamically learns topology features and improves GCN's performance. This model contains more parameters due to the addition of modules, which increases computational needs. Except for the ViT-based model, none of the models above provide interpretability, and all of them have problems doing HAR for complex activities.

## B. Interpretability

Interpretability has a significant impact on whether a user trusts any deep learning model. It is very useful in critical situations such as healthcare diagnostics and surveillance, among others. As we discussed previously, models often lack interpretability, which is a huge problem. Tempel et al. performed a comparative study of Grad-CAM and SHAP and concluded that SHAP provides the "why" behind decisions, while Grad-CAM gives the "where" [8]. They demonstrate that SHAP and Grad-CAM should be viewed as complementary tools rather than competing methods, as Grad-CAM delivers fast spatial explanations while SHAP provides detailed input feature attribution. Lamakaal et al used SHAP to incorporate interpretability in their TinyMSLR model as they intended to reduce model load and minimize parameters [3]. Roy et al, on the other hand, created an interactive visual interface that allows users to import videos, ask questions, and examine model output with explanations [7]. Despite the fact that they evaluate human participants and have acquired high levels of confidence, their model fails to perform in terms of precision and recall.

Grad-CAM has been extensively used for interpretability in image classification and in video- based HAR, however it does not account for the temporal features in videos. Saha et al. do spatiotemporal concept testing in Video-TCAV and find that the model correlates well with how the human brain perceives motion with visual features being processed in initial layers and motion being handled in later layers [6]. They also found that temporal concepts dominate over static concepts in most layers of the model. Fernandez et al. move past conceptualization, attempting to adapt saliency maps and Grad-CAM from 2D to 3D by accounting for temporal dimension when computing GAP gradients [9]. The drawback of these is that they were only tested on a single architecture or dataset.

## C. Attention Mechanism

The problem of occlusion, distraction from background and motion pattern variability that restricts model performance is brought up again, and attention mechanisms can be used to improve model accuracy. We can see hybrid models being embedded with attention mechanisms improves their performance while being interpretable. Manoharan et al. extract spatial and temporal features using an Inception V3 model and Bidirectional Gated Recurrent Units (Bi-GRU) incorporated with an attention mechanism that improves HAR [4]. They replace the tanh function in Bi-GRU with ReLU, which increases accuracy on benchmark datasets like UCF101 (98.13%) and HMDB51 (81.45%). Saoudi et al. employ I3D and Bi-LSTM with a self-attention mechanism [13]. They achieve an F1 score of 94.76on UCF101 and 95.8 on HMDB51. These models are still unable to recognize complex motions or crowded backgrounds, and future development will involve implementing the same using Transformers.

Wu et al. proposes a framework with two modules one focusing on spatial and other on temporal features integrated into a GCN [10]. They apply self-attention to improve the performance of GCN and obtain higher accuracies than other state-of-art methods. However, there is still room for improvement in terms of accuracy. Li et al. introduces Attention Guided Concept Model (AGCM), which accounts for temporal and contextual information and addresses the challenge of multimodal conceptual learning [12]. The authors say the generalizability of the model is an interesting avenue for further investigation.

## D. Literature Summary

Through this literature review, we identified major concerns for researchers in the field of HAR in video sequences, including action recognition in occluded settings, cluttered backgrounds, dimly lit videos and videos with complex actions; the accuracy-interpretability trade-off; and high computational needs and costs. We see that hybrid models outperform their standalone counterparts, and when combined with attention mechanism, the performance improves even further. We also observe a progressive transition from various types of neural networks such as CNN, RNN, and GCN to Transformers. XAI techniques are also being investigated to develop something that can help explain temporal features. Furthermore, researchers are working to make their models intrinsically interpretable, as user trust is critical for these models to be used in the real world.

# 3. SYSTEM ARCHITECTURE

The system is designed around a two-stream pipeline that decomposes an input video into RGB frames and their corresponding optical flow representations, extracts spatial and motion features from each stream independently using a pre-trained Inception V3 network, compresses these features using Global Average Pooling (GAP), fuses the two streams, and then models the temporal evolution of the fused representation using a Bidirectional Gated Recurrent Unit (Bi- GRU) strengthened with an attention mechanism. A distinguishing aspect of the proposed system is that it is trained and evaluated on partial sequences, so that the network learns to anticipate the underlying action before it is fully performed. The system is trained and validated on a mini version of the UCF Crime dataset, constructed by selecting a reduced subset of action classes and video samples to keep the computational requirements of the study manageable.

## A. Data Pre-Processing

The data pre-processing stage converts raw video files from the mini-UCF Crime dataset into standardized, ordered RGB and optical flow frame sequences suitable for the two Inception V3 input branches. Each video is opened using OpenCV's video-capture interface and read frame by frame, producing an ordered collection of RGB frame images that preserves the original temporal sequence of the action.

Optical flow describes how objects move between two consecutive frames of a video through a two-dimensional vector field with horizontal displacement $u$ and vertical displacement $v$. The dense optical flow algorithm computes this $(u, v)$ vector field between every pair of temporally consecutive RGB frames, so that a video with $N$ extracted RGB frames yields $N-1$ optical flow fields.

The direction and magnitude are encoded into the Hue-Saturation-Value (HSV) color space and converted to the RGB color space to produce the color-coded optical flow frame supplied to the optical-flow Inception V3 branch.

During training, every RGB frame and corresponding optical flow frame is extracted, while during validation and testing, the frames inputted are varied, restricting the temporal information available to the network and supporting early action detection. Every extracted RGB frame and every computed optical flow frame is resized to a fixed

resolution of 299×299 pixels with three channels to match the input requirements of the Inception V3 backbones.

## B. Dual-Stream Spatial and Motion Feature Extraction

The RGB stream and the optical flow stream are each processed by their own dedicated Inception V3 network, both following the same underlying architecture but trained with independent weights so that every stream can specialize to its own modality. The Inception V3 backbone is well suited to this role because it combines several parallel convolution filter branches of different kernel sizes with pooling layers, allowing it to capture visual patterns at multiple scales within a single frame. Processing the two modalities independently in this way ensures that neither stream is forced to compromise between representing appearance and representing motion, and it keeps the two sources of information clearly separated until a later fusion stage.

## C. Global Average Pooling (GAP)

Rather than flattening the final feature maps produced by each Inception V3 backbone and feeding them into a dense layer, which introduces a large number of trainable parameters and increases the risk of overfitting, the proposed system applies Global Average Pooling to the final feature map of both streams. GAP condenses each channel of a feature map into a single representative value by averaging across its spatial extent, so that each frame — from either stream — is reduced to one compact feature vector rather than a high-dimensional feature map. Once the RGB descriptor and the optical flow descriptor have been obtained for a given frame, and it is at this point that the two streams merge into one unified sequence.

## D. Temporal Modeling

During training, the unified sequence is divided into temporal windows of 16 time steps, with a stride of 8 during evaluation. Each window is passed to the temporal modeling network.

Initially, a masking layer identifies zero-padded time steps, followed by dropout regularization to reduce overfitting. A fully connected Dense layer with 256 neurons and ReLU activation projects the input features into a lower-dimensional representation. The resulting sequence is processed by a Bidirectional Gated Recurrent Unit (Bi-GRU) with 64 units in each direction. The forward GRU processes the sequence from beginning to end, while the backward GRU processes it in reverse order. Their hidden states are concatenated to produce a 128-dimensional temporal representation at each time step.

The temporal attention mechanism learns the relative importance of each time step in the Bi-GRU output. In the implementation, the attention layer uses 32 hidden units and three learnable parameters: a weight matrix (W), a bias vector (b), and a scoring vector (v).

## E. Algorithm

Input: UCF Crime dataset containing Normal, Road Accidents, and Robbery videos.

Output: Predicted action class

**Begin**

1. Load the video dataset and select the required classes.
2. Split the dataset into training, validation, and testing sets.
3. Extract 64 frames from each video and resize them to 299 x 299.
4. Compute optical flow between consecutive frames to capture motion information.
5. Extract features from RGB frames and optical flow frames using pretrained InceptionV3 networks.
6. Apply Global Average Pooling to obtain feature vectors from both streams.
7. Concatenate the RGB and optical flow feature vectors.
8. Pass the feature sequence through a Dense layer.
9. Pass the sequence through the forward and backward GRU layers.
10. Concatenate the forward and backward hidden states to obtain temporal features.
11. Apply Temporal Attention to assign weights to important time steps.
12. Pass the attention-weighted representation through the Softmax classifier.
13. Perform five-fold cross-validation and calculate accuracy, precision, recall, F1-score, and confusion matrices.

**End**

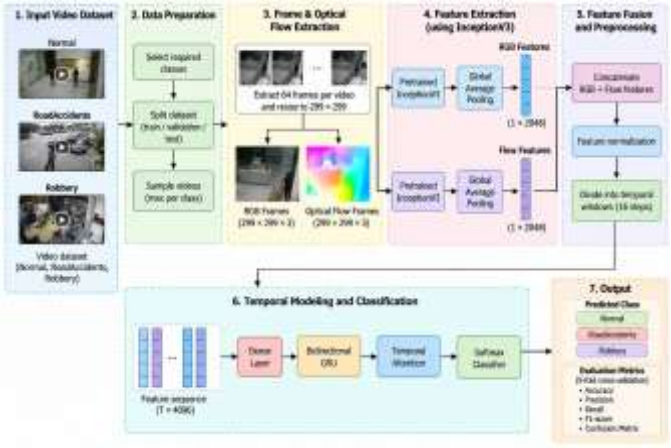



**Fig -1**: System architecture diagram

## F. Hardware Specification

**TABLE 1:** HARDWARE SPECS

| Component | Specification |
|---|---|
| GPU | NVIDIA Tesla T4, 15 GB (15,360 MiB) memory, driver version 580.82.07 |
| CPU | Intel Xeon @ 2.00 GHz, 2 virtual cores |
| RAM | 12 GB |

## G. Software Specification

**TABLE 2:** SOFTWARE SPECS

| Software | Version |
|---|---|
| OS | Linux 6.6.122+ (x86_64, glibc 2.39) |
| Programming language | Python 3.13.15 |
| Open CV | 4.14.0 |

# 6. RESULT & DISCUSSION

## A. Input Dataset

The UCF-Crime dataset is a large-scale dataset of 128 hours of footage. It consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies.

The minimized set used included 250 videos from the classes Road Accidents, Robbery and Normal.

## B. Result of Pre-Processing

The videos from the selected classes were converted into rgb and optical flow frames of dimensions 299 x 299 x 3.




**Fig -2**: Example frame from Road Accidents class




**Fig -3**: Example frame from Robbery class




**Fig -4**: Example frame from Normal class

## C. Result of k-fold cross validation

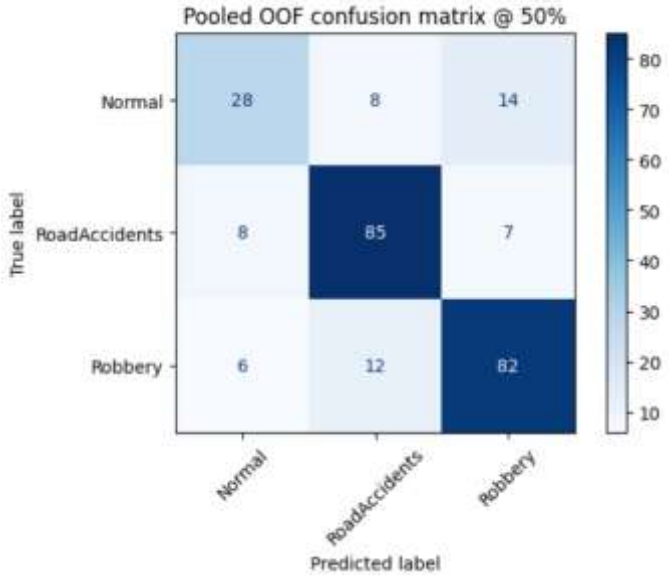
**TABLE 3:** FIVE-FOLD ACCURACY

| Fold | Accuracy at 50% observation | Accuracy at 100% observation |
|---|---|---|
| 1 | 0.76 | 0.74 |
| 2 | 0.8 | 0.78 |
| 3 | 0.72 | 0.76 |
| 4 | 0.82 | 0.8 |
| 5 | 0.8 | 0.82 |

The five-fold cross-validation results show that the proposed model achieves a mean accuracy of 78% at both 50% and 100% observation levels. At 50% observation, the model achieves a macro F1-score of 74.4%, slightly higher than the 73.9% obtained using the complete video sequence. These results indicate that the model can identify discriminative visual and motion patterns from partial video sequences, supporting its potential for early anomaly anticipation in surveillance videos.

**TABLE 4**: MEAN OF ACCURACY, PRECISION, RECALL AND F1-SCORE

| Evaluation metric | 50% observation | 100% observation |
|---|---|---|
| Mean accuracy | 0.78 | 0.78 |
| Mean macro precision | 0.75 | 0.76 |
| Mean macro recall | 0.74 | 0.74 |
| Mean macro F1-score | 0.74 | 0.74 |
| Mean weighted F1-score | 0.77 | 0.77 |




## D. Confusion Matrix

**Fig -5**: Pooled Confusion Matrix for 50% observation window

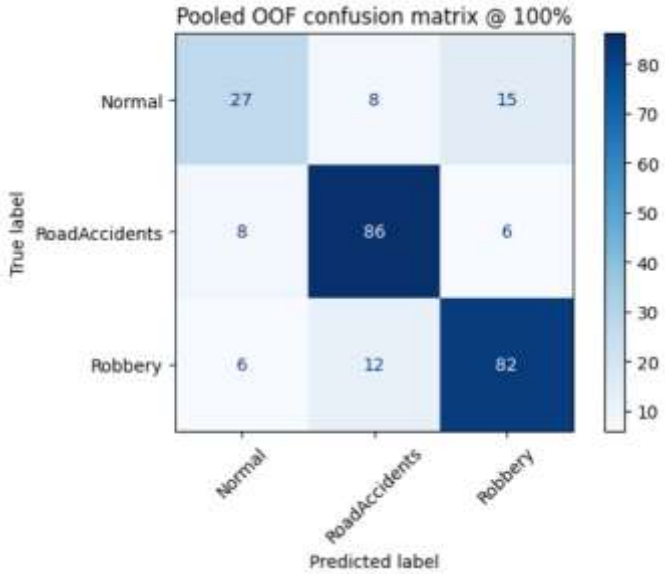



**Fig -6**: Pooled Confusion Matrix for 100% observation window

Overall, the model demonstrates relatively consistent performance across both observation levels, particularly for Road Accidents and

Robbery. The results indicate that useful discriminative features can be extracted from partial video sequences, supporting the potential of the proposed approach for early anomaly anticipation in surveillance videos. However, the lower performance for the Normal class highlights the need for further improvements in distinguishing normal and anomalous behaviour.

### E. Evaluation of variable temporal observation

**TABLE 5:** MEAN ACCURACY ACROSS K-FOLD

| Observed % | Mean accuracy | Standard deviation |
|---|---|---|
| 10 | 0.74 | 0.04 |
| 20 | 0.74 | 0.05 |
| 30 | 0.76 | 0.06 |
| 40 | 0.77 | 0.05 |
| 50 | 0.78 | 0.04 |
| 75 | 0.8 | 0.03 |
| 100 | 0.78 | 0.03 |

The mean accuracy increased from 73.6% at 10% observation to 78.0% at 50% observation, indicating that the availability of additional temporal information generally improves classification performance. The highest mean accuracy of 79.6% was achieved at 75% observation. The standard deviation across folds ranged from 2.8% to 6.4%, with the lowest variation observed at 100% observation and the highest at 30% observation. This indicates that performance was more consistent across folds at higher observation levels.

## 7. FUTURE SCOPE

Increased computational resources would allow for experimenting with more advanced backbone architectures to improve feature extraction, classification accuracy, and generalization. This would also allow us to train the model on the complete UCF Crime dataset. Future work will focus on ensuring that the model predicts an imminent anomaly before the anomalous event actually occurs. This includes training and evaluating the model with pre-onset video segments, as well as monitoring the prediction lead time, to establish how early and reliably anomalies can be anticipated. This also has direct practical uses in public safety and surveillance.

## 8. CONCLUSION

On developing a deep learning-based approach for recognizing anomalies in surveillance videos under variable temporal observation.

The reported five-fold cross-validation results show a mean accuracy of 78% at both 50% and 100% video observation. At 50% observation, the model achieves a macro F1-score of 74.4%, indicating that it can classify the selected action categories using partial video information. The evaluation across different observation levels also shows that the highest mean accuracy, 79.6%, is obtained at 75% observation. These results demonstrate the potential of the proposed architecture for partial-video action recognition.

However, the model has limitations. Its evaluation is based on a reduced dataset containing three classes: Normal, Road Accidents and Robbery. The limited diversity of the dataset may affect the model's ability to generalize to different surveillance environments. Furthermore, recognizing an action from a partial video does not necessarily establish that the model can predict an anomaly before the event begins.

## REFERENCES

[1] R. E. Sri, Y. E. Mulyanto, P. I. K. Eddy, and P. M. Hery, "Human Activity Classification Using Deep Learning Based on 3D Motion Feature," Machine Learning with Applications, vol. 12, p. 100461, June 2023, doi: 10.1016/j.mlwa.2023.100461.

[2] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–10, Apr. 2022, doi: 10.1155/2022/3454167.

[3] I. Lamaakal, C. Yahyati, Y. Maleh, K. El Makkaoui, and I. Ouahbi, "An Explainable Hybrid CNN-transformer Model for Sign Language Recognition on Edge Devices Using Adaptive Fusion and Knowledge Distillation," Scientific reports, vol. 16, no. 1, p. 7143, Mar. 2026, doi: 10.1038/s41598-026-38478-8.

[4] J. Manoharan and Yuvaraj Sivagnanam, "A Novel Human Action Recognition Model by Grad-CAM Visualization With Multi-Level Feature Extraction Using Global Average Pooling With Sequence Modeling by Bidirectional Gated Recurrent Units," International Journal of Computational Intelligence Systems, vol. 18, no. 1, May 2025, doi: 10.1007/s44196-025-00848-x.

[5] H. Wang, Y. Wang, S. Yan, X. Du, Y. Gao, and H. Liu, "Merge-and-Split Graph Convolutional Network for Skeleton-Based Interaction Recognition," Cyborg and Bionic Systems, vol. 5, Jan. 2024, doi: 10.34133/cbsystems.0102.

[6] A. Saha, S. Gupta, S. K. Ankireddy, K. Chahine, and J. Ghosh, "Exploring Explainability in Video Action Recognition," arXiv.org, Apr. 13, 2024. https://arxiv.org/abs/2404.09067 .

[7] C. Roy et al., "Explainable Activity Recognition in Videos Using Deep Learning and Tractable Probabilistic Models," ACM Transactions on Interactive Intelligent Systems, vol. 13, no. 4, pp. 1–32, Dec. 2023, doi: 10.1145/3626961.

[8] F. Tempel, D. Groos, E. A. F. Ihlen, L. Adde, and I. Strümke, "Choose Your Explanation: A Comparison of SHAP and Grad-CAM in Human Activity Recognition," Applied Intelligence, vol. 55, no. 17, Nov. 2025, doi: 10.1007/s10489-025-06968-3.

[9] J. Garcia-Torres Fernandez, "Interpretability in Video-Based Human Action Recognition: Saliency Maps and GradCAM in 3D Convolutional Neural Networks," Nordic Machine Intelligence, vol. 4, no. 2, pp. 19–25, Aug. 2024, doi: 10.5617/nmi.10587.

[10] L. Wu, C. Zhang, and Y. Zou, "SpatioTemporal Focus for Skeleton-Based Action Recognition," Pattern Recognition, vol. 136, p. 109231, Apr. 2023, doi: 10.1016/j.patcog.2022.109231.

[11] Y. Gao, X. Duan, and Q. Dai, "Skeleton-Based Action Recognition Using Graph Convolutional Network With Pose Correction and Channel Topology Refinement," Computers, Materials and Continua, vol. 83, no. 1, pp. 701–718, Mar. 2025, doi: 10.32604/cmc.2025.060137.

[12] X. Li and M. Mahmoud, "Interpretable Concept-Based Deep Learning Framework for Multimodal Human Behavior Modeling," Feb. 14, 2025. Accessed: Oct. 16, 2026. [Online]. Available: https://arxiv.org/abs/2502.10145v1

[13] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, "Advancing Human Action Recognition: A Hybrid Approach Using Attention-Based LSTM and 3D CNN," Scientific African, vol. 21, p. e01796, Sept. 2023, doi: 10.1016/j.sciaf.2023.e01796.