International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I10P122

Volume 74 Issue 10, 308-318, October 2026
© 2026 Seventh Sense Research Group®

Original Article

Explainable Human Action Intelligence for Reliable and Transparent Video-based Recognition

Priyanka Chaudhari[1], Geetanjali Kale[2]

[1, 2]*SCTR's Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, India.*

[1]*Corresponding Author : priyankasavadekar10@gmail.com*

Received: 20 April 2026 Revised: 17 August 2026 Accepted: 26 August 2026 Published: 27 October 2026

Abstract *- The human action recognition discipline has been boosted by the swift developments in Artificial Intelligence to a substantial degree, making it a pillar of the current computer vision studies. The current research contends with essential needs in society, since it uses a self-acquisition paradigm in order to enhance the understanding of intricate physical exercise routines, hence, enabling independent studies. The modern deep-learning models are more likely to promote predictive performance at the loss of interpretability, reliability, and trustworthiness. Their shortcomings are a hindrance in real world application where accountability and open-mindedness are essential. Explainable Human Action Intelligence (EHAI) framework, which seamlessly integrates state-of-the-art deep-feature extraction with thoroughly based explainability modules, which will contribute to putting trust in artificial intelligence. The proposed framework focuses on improving both the recognition accuracy and interpretability of human action recognition. ResNet-50 is refined using the UCF101 dataset, which comprises 101 human action categories. Gradient weighted Class Activation Mapping approach (Grad-CAM) is incorporated to localize and visualize salient regions within video frames, providing insight into the visual evidence influencing the model's predictions. Transfer learning is employed to adapt ResNet-50 to the target action recognition task, resulting in an accuracy of 95.4%. The explainable AI methods allow generating visual explanations that underline the areas that are the most appropriate to the model predictions and promote transparency, reliability, and trust. As a result, the state-of-the art framework is not only better in current scans of human action recognition, but also improves model accountability significantly, which additionally promotes the clarity of AI-based resolutions.*

Keywords *- Explainable Artificial Intelligence (XAI), Human Action Recognition, Deep learning, Model explainability, Trustworthy AI.*

1. Introduction

In cognitive science, it is often believed that human vision deduces objects or goals in particular scenes out of a number of characteristics like form, color, size and other low-level visual hints. Attribute labelling determines certain "attributes" on human bodies, hence the reason why the CNN model might be. It is a well-established belief in mind science that when a human observes a scene, it tries to detect objects or findings in that context by looking at numerous low-level visual features, including shape, colour, size and other primitive elements. The idea of attribute labelling, assigning certain attributes to human bodies, is what the purpose of the Convolutional Neural Network (CNN) model is based on, as this type of model can automatically acquire and classify these attributes based on visual images. The Human Action Recognition (HAR) method can be a way by which users learn the process involved in an activity. In computer vision, human action recognition is significantly enhancing the automation and effectiveness of numerous applications like video indexing, sports analytics, surveillance systems, posture detections, human-computer interaction. Enhanced HAR leads to improved security measures, posture detection, effective dance training, improved gym working out correction, and better performance analysis in sports. This study seeks to enhance the performance of human action detection and recognition with Explainable Artificial Intelligence (XAI) in order to demonstration the model's robustness. Traditional methods for HAR mainly depended on heuristically crafted features together with traditional machine learning algorithms, necessitated domain expertise and had limited generalization capabilities across several datasets. HAR has been completely changed since deep learning came into play, where models are able to automatically understand hierarchical representations from raw video pixels, making them more accurate and robust than ever before. For example, take intelligent tutoring systems, which are developed to help students in learning activities by identifying the actions of involved students. Human action recognition has also assisted robotics. Industrial robots, for instance, that provide assistance to workers have the ability to emulate humans' spontaneous

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

movements, thus refining information flow throughout the work process. Over recent years, technology has supped deeply into day-to-day life and, mostly in work environments. The capacity to anticipate the actions of latent human mistakes in work processes is increasingly a consideration in tasks. Human Action Recognition (HAR) can be a method that enriches the upgrading process of users in an activity, rejecting doubtful decisions. This study aims to enhance human action recognition through the integration of explainable AI to prove the model's robustness in the real world. The previous methods of human action recognition depended on the usage of handcrafts and simple machine-learning methods that needed to be trained with a great deficiency of domain knowledge and could fail to transfer across a variety of datasets. With the introduction of the deep learning concept, the dynamism of this scene has completely changed, and systems can now automatically learn hierarchical representations of raw video, which significantly increases the recognition accuracy. Convolutional neural networks, among the many different models of Deep-Learning, have enjoyed considerable success in action recognition, of which they outperform their predecessors in an extensive range of benchmarks.

The Explainable Artificial Intelligence (XAI) brings new possibilities to enhance the interpretability, explainability and trustworthiness of AI systems. This study emphases on enhancing human action recognition using Explainable AI along with visualization. The system helps to provide detailed explanations for highly precise XAI systems on human actions by upsurging transparency and understanding the process of decision making. UCF101 Human action dataset is used for action recognition and to generate the explanations. The generated explanations bridge the gap between system's decision about recognized actions and actual actions. Significant features are identified which are contributing to improve the accuracy of human action recognition. This system aims to shed light on the Explainable AI model. It also improves utilization of such Explainable AI classification models in the scenarios of real world. Recent research has focused on improving the explainability of human action recognition using Grad-CAM. Ensuring that models' decision-making processes in the actual world give due diligence is an important branch of machine learning, and explainable AI empowers users to be more careful. In addition, HAR tries to perceive human actions using data from RGB-D cameras in order to facilitate human-machine interactions.

The Grad-CAM explainable method generates attention maps, which generate classifications that are present in the action recognition model, allowing understanding where the classifiers are focusing on the action performed in the image. These maps can also show how the perturbations have acted on the action recognition system, offering a qualitative comparison between the strength of the results. Over the last decade, there has been a widespread consensus that the vast majority of AI models, such as Neural Networks (NNs), are classified as black boxes due to a lack of transparency about how and why they make a judgment or prediction. This leads to the creation of numerous publications on Understanding the Reasoning of Neural Networks for Human-Machine Collaboration [19]. In this study, an Explainable AI method is investigated for human action recognition using a deep convolutional neural network model. The Grad-CAM approach achieves the explanation, and it is used in an action recognition system with action classification, making it a more efficient alternative to traditional hard annotation. For example, intelligent tutoring systems have been built to assist pupils in learning. Intelligent tutoring systems, for example, have been intended to help students with learning tasks by identifying the exact activities in which they are participating. Human machine interaction has also profited from human action recognition [7, 17].

For example, machines that aid workers in industrial operations might emulate spontaneous movements produced by humans, enhancing the flow of communication in the workplace. They also act on mental health impairments by emulating a teacher's movements to help the student treat the disability. Recognizing and comprehending human activities is a necessary skill for any intelligent agent that interacts or collaborates with humans. This challenge is especially important in the field of interaction between humans and machines, which seeks to develop efficient, natural, and intuitive methods of interacting with machines that are simple and intuitive for end users. These end users include adults, the elderly, and children with varied levels of technological familiarity, and AI-based ideas are integrated in a simple and visible manner. The explainable system that delivers three types of explanations a) Video Explanations: When the system highlights (potentially many) portions of the video in which the Action is recognized. If no comparable activity is detected in the event of a "no answer," the proposed system should display the most likely activity in the video. b) Ranked explanations: predict the activity in object, location, action in a video. C) Most likely entities: The system aims to identify and present the actions with the highest predicted likelihood, objects, and places, as well as their likelihood of being related to the query [20, 26].

A deep residual network with 50 layers of Residual Network 50 (ResNet50) architecture introduced. has gained popularity for complex tasks like object detection, prediction, image recognition, and classification. The ResNet50 model focuses on the problem of vanishing gradients in deep neural networks with integration of residual connections without degrading the model's performance. This is the motivation to use the fine-tune ResNet50 model for human action recognition in videos. The ResNet50 deep model allows to identify of complex features and hierarchical depictions, which help to accurately recognize complex human actions in video data. The residual block, which mitigates the vanishing

gradient problem. This Residual block makes the ResNet50 model robust and apprehends fine-grained details in human action. The Transfer learning influences action recognition and classifications using ResNet50, which is also effective for the UCF101 dataset [11, 12].

1.1. Need for Model Explainability

The deep neural models need to ensure interpretability and transparency while achieving high accuracy in human action recognition. Human Action recognition system is not only ensuring model's the high accuracy but also makes model transplant through generating the explanation on models' classification decision. Explainability in AI helps to generate trust and accountability about models among the end users and stakeholders. Explainability helps to gain user confidence in critical applications like smart surveillance, medical imaging, healthcare applications, the entertainment industry, anomaly detection, elderly care, and rehabilitation [6]. The validation and verification of the model's behavior are ensured by Explainable Artificial Intelligence. Explanation is generated based on the relevant features rather than the irrelevant correlations [24]. Visualization techniques enable us to pinpoint the question of 'Why model make the particular decision?'. This is invaluable for error detection, analysis, and debugging of model performance. The ethical and Legal compliance is endorsing meeting requirements with the help of Explainable Artificial Intelligence [1, 9, 28]. Explainable AI facilitates human-machine interaction by providing visions into the model's reasoning process [8, 18].

Although deep learning-based Human Action Recognition (HAR) methods have achieved substantial improvements in classification performance, a significant research gap remains in understanding and interpreting the decisions produced by these models. Most existing HAR approaches primarily focus on improving recognition accuracy through deeper architectures and enhanced feature representations, while comparatively limited attention is given to explaining the visual evidence responsible for individual predictions. Consequently, high recognition accuracy alone does not establish whether a model has learned action-relevant characteristics or has relied on background objects, scene context, or other irrelevant visual cues. Furthermore, many existing studies evaluate HAR models predominantly using quantitative performance measures, with limited integration of interpretable visual evidence into the recognition and evaluation process.

To address this research gap, the present study develops and systematically evaluates an explainable HAR framework that integrates ResNet50-based action recognition with Gradient-weighted Class Activation Mapping (Grad-CAM). ResNet50 is employed to learn discriminative visual representations for action classification, whereas Grad-CAM provides class-specific activation maps that reveal the spatial regions contributing to individual predictions. The contribution of the proposed framework lies in complementing quantitative recognition performance with visual interpretability, thereby enabling a more transparent examination of the model's decision-making process. Experimental results demonstrate a classification accuracy of 95.4%, while the Grad-CAM visualizations provide additional insight into the action-relevant regions influencing the predictions. Thus, the proposed framework addresses the identified gap by considering both recognition effectiveness and interpretability rather than evaluating HAR solely from the perspective of classification performance.

The proposed state-of-the-art system combines the vigorous feature extraction abilities of ResNet-50 with Grad-CAM visualization methods to recognize human actions that are both accurate and explainable. The system enhances transparency and provides important visions of the model's decision by explanations through visualizing the regions of video in the video frames. This system not only improves the model performance but also endorses trust and liability in human action recognition with classification systems. The foremost goals of the paper are:

- Demonstrate the transfer learning of fine-tune deep learning architecture ResNet-50 model on the UCF-101 dataset, with accuracy in recognizing a wide range of human actions.
- Employing the Grad-CAM (Gradient weighted Class Activation Mapping) method allows for the visualization of critical regions within video frames during action recognition.
- The system bridges gap between the performance of the human action recognition model and user interpretability, which arise the need of explainability by promoting trust and transparency.
- This study contributes to the emerging area of Explainable Artificial Intelligence by performing practically.

2. Related Work

The recognition of Human Action Recognition (HAR) has had a very strong development on the development of deep learning architectures, however, the lack of transparency in deep learning models has directed to apprehensions on interpretability, trust, and trustworthiness of the models in the decision-making. Current surveys on Explainable Artificial Intelligence (XAI) point out that the attainment of trustworthy AI through a carefully designed combination of interpretability mechanisms, validation of transparency and domain-specific evaluation strategies is valued [1, 6, 7]. These publications highlight the fact that explainability will no longer be an auxiliary quality, but it will form part of the foundational basis of implementing HAR models in sensitive domains like surveillance, medical care, monitoring of ultra-vires through UAV, and smart environments.

The superficiality of convolutional and residual networks still prevails in HAR because of the representational ability of deep learning. Addition of residual learning [11, 12] and treatment of individual image and temporal motion cues led to a bigger enhancement in recognition performance [14-16]. Most of the more recent literature explores dynamic motion representation and synthetic training approaches to avoid the limitations of a limitation of annotated video data in the real world [2, 10]. Complementary attempts look at spatial-temporal contrastive learning as well as pyramid-based video representation learning [22, 31], and will reveal a transition to motion-aware embedding learning, which has the ability to learn long-assortment dependencies that are imperative to action understanding.

In this fast-developing environment, the importance of explainability is even stronger. The need to determine the relationship between model characteristics and human comprehensible characteristics is supported by XAI studies, which also require that predictions made by the model be in line with domain logic as well as user expectations [8, 18, 19]. Interpretability is crucial in the context of HAR since human activity is frequently characterized by delicate, overlapping, or ambiguous motion patterns that put a strain on the consistency of a model as well as its interpretation by a human. The explicit literature literature that covers video-based explainability shows that the generated explanations should be both semantically correct and they should also be true to the internal processes of the model [20, 26].

The combination of Grad-CAM and extensions constitute a major foundation of visual explanation in deep networks. Grad-CAM Grad-CAM formulation was used as the basis to introduce gradient-based localization to indicate discriminative regions [28], and Grad-CAM+ added even more spatial localization to multi-instance cases. These methods have been popular in vision problems and are increasingly used in HAR, where highlighting motion-relevant regions assists in making sense of the time-varying reasoning of CNN and hybrid networks. Other works demonstrate that the relevance of motion and attention to time could also be used to estimate the accuracy of the explanation in a video recognition context [13, 17], which is also crucial in balancing Visual attention with human cognitive perceptions.

Explainability has also been found useful in sensor-based and smart-home HAR applications to diagnose errors in the system, detect behavioral anomalies, and enhance trust with the users. Recent publications have evidenced that adding interpretable modules to HAR pipelines does not lead to a drop in computational efficiency, but helps to improve pipeline transparency [5, 30]. Similarly, adversarial explanation research findings show that model interpretations can be highly misleading due to perturbations, which underscores the significance of robustness-sensitive XAI evaluation [31]. Prior knowledge and state-transition modelling can be used in video action reasoning to provide more causal information about model behaviour [33].

Outside HAR, explainability systems like ExplAIner [29], LIME [24, 25], and global black-box explanation taxonomies [9] can be conceptually grounded and are applicable to human-centric vision systems. New interdisciplinary studies also tie interpretability to more overarching epistemic and scientific objectives in extending XAI to technical transparency into trust and fairness and useful understanding [1, 8]. A similar pattern is observed in medical explainability experiments where Grad-CAM was applied to the critical diagnostic tasks [34], resulting in visual interpretability interpolating the model outputs and expert justification, an ultimate goal also driving explainable HAR.

Altogether, the literature exposes the obvious trend: the opaque high-performance models are being phased out in favor of understandable and reliable systems with the potential to defend the predictions. Nonetheless, the combination of Grad-CAM-based visualization with a time-related reasoning system, multi-modal motion signal, and real-time HAR systems is yet to be made. To address this gap, there is a need to employ processes that bring together precision and truth so as to have the capacity of offering finer details of spatial-temporal clarifications that enable determination of the significance of features across frames, modalities and decision routes.

Although the domain of deep learning-based approach HAR has made major innovations, and Explainable AI (XAI) has become more mature, there are still several important gaps in research. To begin with, the present HAR models have a high level of accuracy, yet they remain opaque black-box systems that restrict their use in safety-critical areas, as overwhelming evidence points to the importance of being a trusted and verifiable system [1, 6, 7, 19]. The majority of existing XAI systems were initially designed to solve the problem of single image classification, and do not explicitly express the temporal relationships present in human activity, which leads to explanations not properly representing the natural motivations of motion or temporal aptitude [13, 20, 26].

Second, Grad-CAM and its improved versions like Grad-CAM++ [3, 28] give useful visuospatial responses, but they are still weak in their capacity to represent dynamic frame-to-frame significance to video sequences. The existing research demonstrates that these visualizations tend to highlight solid discriminative areas instead of highlighting timely subtleties, which are the main focus of HAR activity. The lack of a consistent framework of spatial-temporal interpretability limits their capabilities of evaluating the manner in which models assimilate motion data over different timeframes and in different complexities [2, 10, 23]. Third, the current

explainability techniques do not have sound evaluation systems that can determine the fidelity, permanence, and reliability of explanations generated. Even though multiple studies demonstrate that it is necessary to validate the rightness and truthfulness of explanations [20, 26, 32], there are no widely used metrics to ensure consistency in benchmarking, and interpretability claims are mostly qualitative and subjective, as well as not reproducible. Furthermore, the adversarial analysis demonstrates that the explanations are easy to intercept, which points at the fragility of the existing XAI solutions and the necessity to seek a way to bridge the gap between explanations and resilience [32]. Fourth, the gap between explainability methods and the real-life implementation of HAR is still insufficient. Surveillance and UAV surveillance. Lightweight Condition surveillance and UAV-based surveillance models [14, 21, 30] tend to compromise interpretability in favor of speed, and smart-home HAR systems [5] struggle to incorporate explanations within ongoing multi-modes of data. Besides, contrastive and synthetic learning methods [10, 22, 31] have enhanced model generalisation, but do not have analogous decipherable access to the way such depictions are utilised to result in ultimate predictions. Recent work in [35] explored attention mechanisms within 3D video classification models to strengthen spatiotemporal feature learning for action recognition. In [36], motion-based representations were investigated for privacy-aware cross-domain action recognition, reducing dependence on appearance information while preserving action-related motion cues. Lastly, in spite of the arguments propose by interdisciplinary researchers about explainability as the means of uniting human thinking with the decisions made by AI [8, 17, 18], present-day HAR research lacks the provision of the complete correspondence between the explanations and the motion cues as perceived by people. It leads to a disconnect between what the domain experts deem important action characteristics and the things highlighted by the model. Thus, it is clear that the techniques of explaining need the combination of spatial saliency, temporal dynamics, and human-consistent attention to ensure reliable and action-taking interpretation.

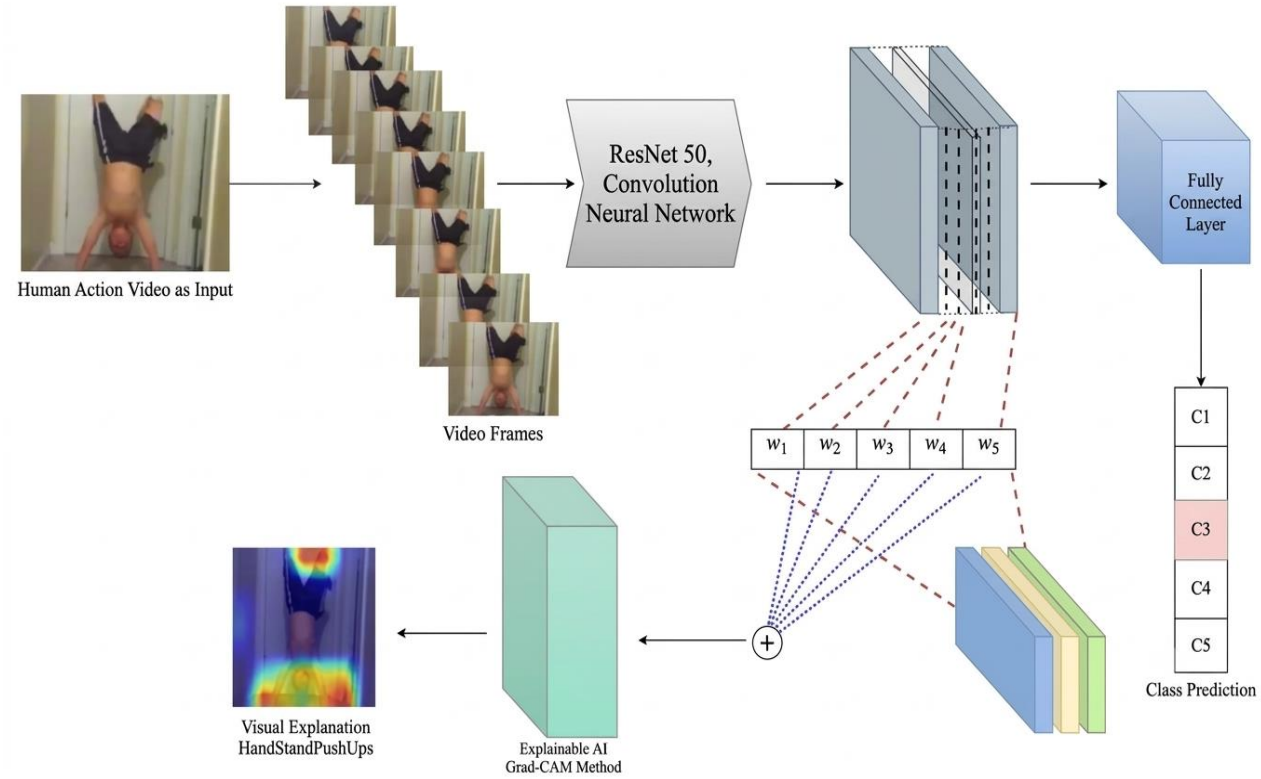

Fig. 1 Proposed methodology

3. Proposed Methodology

The system used a convolutional neural network of finer tuned ResNet50 that is a deep architecture with a residual learning capacity. The research problem is the application of ResNet50 model to the problem of human action recognition, which is quite a daunting assignment. Along with the prominence of high accuracy in recognizing human actions, the system also emphasizes explainability in AI models. The proposed system is employing Gradient weighted Class Activation Mapping approach to see the attention of the human action recognition model focus is on in a video frame. This bilateral approach is based on high-performance requirements and explainability and offers a solid architecture of the improved human action recognition systems.

This is a deep model to action recognition with a dual approach, and there is another with explaining model applied with Grad-CAM method. Fine-tune ResNet-50 model for Human action recognition using ResNet-50 Convolutional Neural Network, make straightforward the guiding of deep networks by easing the declining gradient challenge. ResNet-50 was preferred for its balance of depth and computational competence, achieving it appropriate aimed at complex, critical recognition assignments such as human action recognition.

In this work, the proposed system classifies human action videos of multi classes from UCF 101 human action dataset by employing a pretrained ResNet50 transfer learning model, as shown in Figure 1 and the algorithm.

3.1. ResNet-50 Architecture

ResNet-50 is a 50-layered convolutional neural network. Residual connections are mainly employed in ResNet, which easily sanctions gradients to flow through the network through preparation, thus empowering the training of deep layers [11].

3.1.1. Residual Block

The conventional neural network, each deep layer passes into the succeeding deep layer, but while employing residual blocks, each layer maintains into the next layer and precisely obsessed by the layers about 2–3 phases away.

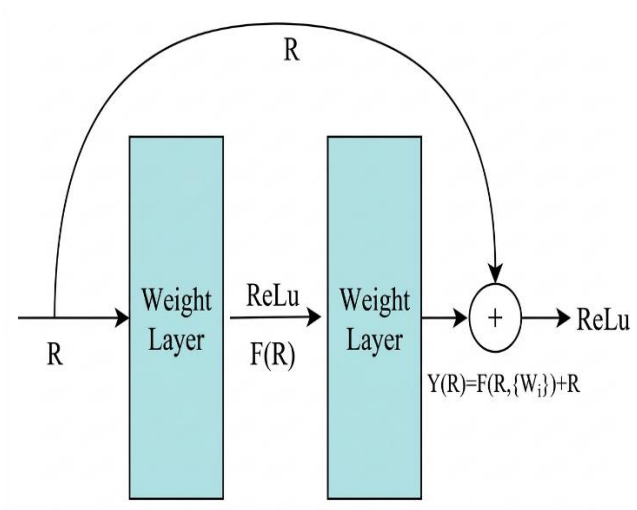

Fig. 2 Residual block

The Input of the residual block is 'R' as expressed in Figure 2. The residual mapping is represented as 'F(R,{Wi})', which typically encompasses a sequence of convolutional layers, batch normalization, and Rectified Linear Unit (ReLU) activations. As illustrated in Figure 2 and Equation (1) Output Y(R) of the residual mapping also consists of the 'z' with providing residual functions with reference to the input.

$$Y(R) = f\ (R, \{Wi\}) + R \qquad (1)$$

Deep convolutional neural network, ResNet-50, consists of 50 layers shown in Figure 3, involving convolutional layers, pooling layers, and fully connected layers.

3.1.2. Convolutional Layer

The set of passes through a filter are employed by a convolutional to the input video frame to generate feature maps. Input video frame is represented by 'I', and 'K' is represented as a convolutional filter. 'p' and 'q' are the coordinates of the convolutional filter size. (i, j) are the organizes of the yield image pixel.

The convolution operations for a single filter can be written as Equation (2):

$$(IK)\ (i,j) = \sum p\ \ \sum q\ \ K(m,n)\ I(i+m, j+n) \qquad (2)$$

3.1.3. Explainable Model using Grad-CAM

Grad-CAM Explainable AI method, which generates the visual explanations on human action recognition and classification from a HAR algorithm by highlighting significant regions in the input video frame.

Gradient-weighted Class Activation Mapping mostly considers gradient information, which is generated by the last layer of convolution layer of the ResNet50 convolution neural network. Grad-CAM is used to visualize any activation in deep neural networks, among that it focuses on explaining decisions taken by Neural networks [3, 28]. Visualizing explanations that consist of a significant part of the video frame, which led CNN to recognize and classify human action. The Grad-CAM method generates heatmaps that represent activation classes on the most contribution frame of a video input. Each activation class is associated with a specific human action class.

The Class Activation Map (CAM) allocates importance to each position (x, y) in the extant convolutional layers shrewd the linear combination of activations, weighted by the comparable output weights designed for the detected class.

Grad-CAM generates visual explanations for classification from a ResNet 50 by highlighting significant regions in the input, most contributing video frames. The steps for Grad-CAM are:

3.1.4. Gradient Calculation

$$\alpha_k^C = \frac{1}{z} \Sigma_i \Sigma_j \frac{\partial y^C}{\partial A_{ij}^k} \qquad (3)$$

Gradient weight for feature map 'k' with respective to each human action class 'C' is expressed as α_k^C. In the given Equation (3) A_{ij}^k is the activation of unit k in layer A at spatial position (i, j). Before softmax, y^C is the score for class C. Number of pixels z in the feature map.

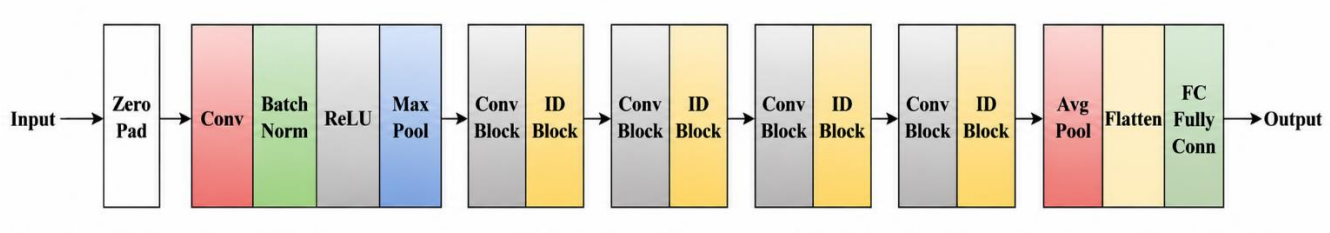

Fig. 3 Structure of ResNet-50

Algorithm 1: Human action identification and recognition using deep residual network

Algorithm: Human Action Identification Recognition using Deep Residual Network

Inputs: $D = \{C_1, C_2, C_3 \ldots\ldots C_n\}$ A Set of Distinct Activity Classes
$A = \{V_1, V_2, V_3 \ldots\ldots V_t\}$ A set of various Video files
$V = \{f_1, f_2, f_3, \ldots\ldots\ldots f_t\}$ A set of Frames
f_t = set of features
Where F_t is t^{th} frame, $F_t \in \mathbb{R}^{m*n*c}$ representing pixel value with height m, width n and colour channel c

Output: V_o = Output video with recognized action

Begin:
while D ! = Ø do
Selection of randomized video V_i from D
while V_i is open do
Iterate through the $V = \{f_1, f_2, f_3, \ldots\ldots\ldots f_t\}$
for each frame f_t in V_{ij} do
Extracted Features $f_t^{S,CNN}$
Transform features into class scores.

$$Fully\ Connected\left(f_t^{S,CNN}\right) = Z$$

$$Z = W_{\mathrm{FC}} \cdot f_t^{S,CNN} + B_{\mathrm{FC}}$$

Z= Output logits (raw scores for action classes).

W_{FC} =Weight matrix (learnable parameters (No of classes X No of Features).

$$W_{\mathrm{FC}(t+1)} = W_{\mathrm{FC}(t)} - \eta \cdot \frac{\partial L}{\partial W_{\mathrm{FC}}}$$

B_{FC} = Bias term (one per class).

$$B_{\mathrm{FC}(t+1)} = B_{\mathrm{FC}(t)} - \eta \cdot \frac{\partial L}{\partial B_{\mathrm{FC}}}$$

$$\eta = Learning\ Rate\ (0.001)$$

$$\mathcal{L} = -\sum_{c=1}^{C} A_c \log\left(\widehat{A_c}\right)$$

Convert the raw scores into Class probabilities
Action classification

$$\hat{A}_{tc} = softmax(Z)$$

$$\hat{A}_{tc} = \frac{e^{Z_i}}{\sum_{j=1}^{C} e^{Z_j}}$$

Z_i = the raw class score for class i

Z_j = All class scores

end
Predicted action class on $(Predicted_{Class})$

$\hat{A}_{tC} = arg \ \ max \ \ \hat{A}_{tC}$ $\quad C\epsilon\{1,2, \dots \dots n\}$

end

end

3.1.5. Heatmap Generation

Grad-CAM Heatmap $L^C_{G\,rad-CAM}$ is generated for class-specific C in eq (4). The positive and significant contributions are attributed to the utility of the Rectified Linear Unit (ReLU) activation function.

$$L^C_{G\,rad-CAM} = Re\,LU\left(\Sigma_k \alpha^C_k A^k\right) \quad (4)$$

The UCF 101 video dataset is used, and splitting the dataset randomly into a training set and testing set (validation). Sets of train videos are pre-processed with feature extraction. Pretrained ResNet50 model weights are used, which are trained on different classes of action videos in the UCF101 dataset. Pre-trained weights with fine-tuning the model ResNet50 are loaded and tested on a test set of video data. Finally, the model recognizes the specific human action among multi classes and achieves acute action recognition accuracy.

For example, if a video of a person performing pull ups from the UCF101 dataset consider as human action video, which is used in a convolutional network of human action, the Grad-CAM visualization can generate a heatmap for the pull ups action class, signifying the extent to which the dissimilar portions of the video frame correspond to a pull-ups human action class. Grad-CAM is utilized to highlight the most influential regions for human action recognition and classification. This aids in understanding the crucial features used in recognizing specific classes.

The visualization of highlighted regions that contribute the most is achieved through a fine-tune ResNet50 model. Such visualization results in enhancing action recognition using convolutional neural networks. The visual explanations generated trust and transparency of the Deep Learning CNN model, the interpretable explanations also help to understand why CNN model made crucial decisions like the specific Human action recognitions and action classification.

4. Results and Discussions

The present investigation was conducted to assess the efficacy of a modified ResNet-50 architecture on the UCF-101 dataset for the purpose of Human Action Recognition (HAR).

Based on the ResNet-50 backbone, which is trained on the ImageNet, which is famous in provisions of efficient feature extraction, it fitted ought of the dataset UCF-101 data, given the particularities of the latter dataset. The videos were divided into separate frames, and the series of frames were consumed to the model. The video frames were also resized to reach the dimensions of 224 x 224 pixels, and then, frames were stripped away to create sequences which would be used to feed into the model.

The system ensures that the pre-trained weights are intact, with the first layers of ResNet-50 were frozen at the training stage and the remaining layers retrained to optimize the model.

Different data augmentation procedures, such as random cropping, flipping, and rotation, were carried out to augment the generalization capabilities of the model. Data was divided into training part (70%), validation (15%) and test (15%) programs to determine the performance of the model.

The scheme operation of the system is presented in Figure 4 by the example of handstand Walking actions, push-ups, punches, and pull-up actions. Action handstand, walking, PushUps, Punch, and PullUps are explained with heatmaps as a result of using Grad-CAM. The model of ResNet-50 that was fine-tuned achieved an accuracy of 95.4% in the examination dataset, which is indicative of its effectiveness in predicting human motions. Both the training and the validation losses achieved a convergence point after 50 epochs; this is a demonstration of consistent learning. Application of the Grad-CAM was introduced to create thermal maps with the focus on identifying which parts of the frames carried the most relevant data in regard to the model projections.

Table 1. Comparative study with proposed approach

Year	HAR Method	Dataset	Accuracy	Explanations
2017 [30]	Spatiotemporal Pyramid Network	UCF101	94%	X
2022 [31]	CNN-based HAR	15-class HAR dataset	71.54%	X
2023 [23]	3D Motion Features + Deep Learning	9-class activity dataset	94.08%	X
2025 [35]	Multi-Head Attention	UCF101	88.98%	X
2026 [36]	Motion-I3D	UCF101	83.0%	X
Proposed Approach	ResNet50 + Grad-CAM	UCF101	95.4%	✓

These graphical representations provided some important information about the decision-making procedure of the model system. An example is in the illustration given, in the Handstand Pushups exercise, Grad-CAM has highlighted the human body, movements of the person, i.e. the concentration of the model on the key elements of actions, hence affirming the proper identification of key elements of actions. The Grad-CAM representations were a means to have a clear view of the focal points in the model and a boost on confidence and validation. Understanding the area of focus of the model helped in identifying the areas that required improvement, including refinement of feature extraction to similar actions. Table 1 compares proposed approach methods. By developing a fine-tuned ResNet-50 model and combining it with Grad-CAM visualization approach better outperformed these methods with an ultimate rate at 95.4% of accuracy, demonstrating the effectiveness and robustness of proposed model. The system is straightforward and honest to the user to comprehend the model interpretability, which aids on generating trust on artificial intelligence models.

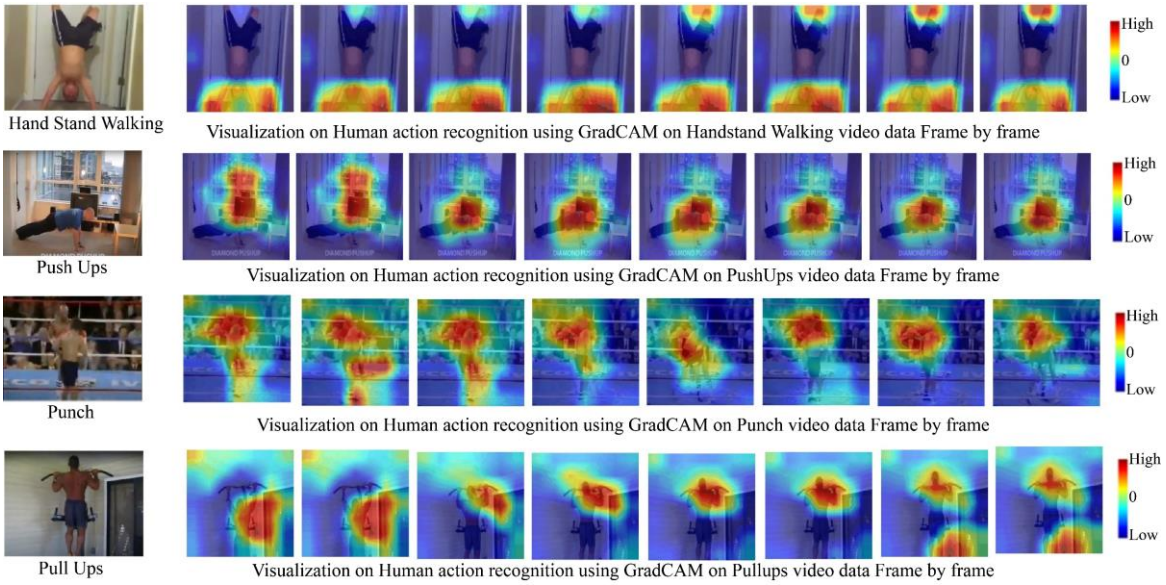

Fig. 4 Visualization using Grad-CAM on recognized human action on videos

5. Conclusion

The paper shows the effectiveness of a fined ResNet-50 architecture in Human Action Recognition (HAR) on the UCF-101 dataset with respectable accuracy of 95.4% per cent. Through visualisation, Gradient-Weighted Class Activation Mapping (Grad-CAM), the proposed system has been able to explain the predictive reasoning behind the model with a particular emphasis on those areas on the input frames that are the most influential to its decision-making process. The proposed system performance can be compared to popular HAR methodologies, and visual analysis makes sure that the network prioritises the essential components of action in the appropriate way. The results, in turn, contribute to the HAR literature by using a well-defined deep-learning backbone, ResNet-50, and Grad-CAM visualisation, hence the increased interpretability. Not only does this confluence help to maintain the impressive 95.4% per cent recognition accuracy, but it also provides a deeper comprehension of how the model works internally, which makes it easier to read and trust. Grad-CAM usage can help understand what exactly in the input data the model is focusing on and is therefore beneficial in overcoming the drawback of opaque models advocated by modern deep-learning models, and contributes to the explanation AI in video action recognition research.

Primarily, it is recommended that further studies explore numerous possibilities to emerging HAR and explainable AI. To better represent temporal feature extraction, convolutional neural networks can be combined with complex temporal architectures like Transformers to better model long-range interdependencies and complex temporal dynamics. In addition, there is more sophistication in data-augmentation approaches and the generation of synthetic data that should be investigated to support model robustness and generalisation. Through these emerging research paths, can increase the accuracy, robustness, and intelligibility of the HAR systems, resulting in the building of more credible and responsible AI models in this field.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Author Contributions

P.C.: investigation, methodology, writing an original draft; G. K.: conceptualization, investigation, reviewing and editing. All authors have read and agreed to the published version of the manuscript.

References

[1] Sajid Ali et al., "Explainable Artificial Intelligence (XAI): What We Know and What is Left to Attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, pp. 1-52, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps, "Dynamic Motion Representation for Human Action Recognition," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1234-1243, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Aditya Chattopadhay et al., "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, pp. 839-847, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ahmet Çınar, Muhammed Yıldırım, and Yeşim Eroğlu, "Classification of Pneumonia Cell Images using Improved ResNet50 Model," *Signal Processing*, vol. 38, no. 1, pp. 165-173, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Devleena Das et al., "Explainable Activity Recognition for Smart Home Systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 2, pp. 1-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Finale Doshi-Velez, and Been Kim, "Towards a Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv*, pp. 1-13, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić, "Explainable Artificial Intelligence: A Survey," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 210-215, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] M. Beatrice Fazi et al., "Beyond Human: Deep Learning, Explainability and Representation," *Theory, Culture and Society*, vol. 38, no. 7-8, pp. 55-77, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Riccardo Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1-42, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Xi Guo et al., "Learning Video Representations of Human Motion from Synthetic Data," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5678-5687, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Kaiming He et al., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. [Google Scholar] [Publisher Link]

[12] Kaiming He et al., "Identity Mappings in Deep Residual Networks," *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, pp. 630-645, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[13] Liam Hile et al., "Explaining Motion Relevance for Activity Recognition in Video Deep Learning Models," *arXiv preprint arXiv*, pp. 1-9, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Narit Hnoohom, Sakorn Mekruksavanich, and Anuchit Jitpattanakul, "An Efficient ResNetSE Architecture for Smoking Activity Recognition from Smartwatch," *Intelligent Automation and Soft Computing*, vol. 35, no. 1, pp. 1245-1259, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Shuaidong Ji, "Multi-Classification of Human Action based on ResNet," *Highlights in Science, Engineering and Technology*, vol. 23, pp. 1-6, 2022. [CrossRef] [Publisher Link]

[16] Yixue Lin et al., "Human Action Recognition Algorithm based on Improved ResNet and Skeletal Keypoints in Single Image," *Mathematical Problems in Engineering*, vol. 2020, no. 1, pp. 1-10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Guoyang Liu et al., "Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models," *Neural Networks*, vol. 177, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Tim Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Saumitra Mishra et al., "A Survey on the Robustness of Feature Importance and Counterfactual Explanations," *arXiv preprint arXiv*, pp. 1-8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mahsan Nourani et al., "Don't Explain without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition," *arXiv preprint arXiv*, pp. 1-26, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Nashwan Adnan Othman, and Ilhan Aydin, "Development of a Novel Lightweight CNN Model for Classification of Human Actions in UAV-Captured Videos," *Drones*, vol. 7, no. 3, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Rui Qian et al., "Spatiotemporal Contrastive Video Representation Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 6964-6974. 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Endang Sri Rahayu et al., "Human Activity Classification using Deep Learning based on 3D Motion Feature," *Machine Learning with Applications*, vol. 12, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should I trust you? Explaining the Predictions of any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[25] Chiradeep Roy et al., "Explainable Activity Recognition in videos: Lessons Learned," *Applied AI Letters*, vol. 2, no. 4, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Avinab Saha et al., "Exploring Explainability in Video Action Recognition," *arXiv preprint arXiv*, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Ramprasaath R. Selvaraju, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017. [Google Scholar] [Publisher Link]

[28] Thilo Spinner et al., "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1064-1074, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[29] Amin Ullah et al., "Efficient Activity Recognition using Lightweight CNN and DS-GRU Network for Surveillance Applications," *Applied Soft Computing*, vol. 103, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Yunbo Wang et al., "Spatiotemporal Pyramid Network for Video Action Recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8987-8996, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[31] Walt Woods, Jack Chen, and Christof Teuscher, "Adversarial Explanations for Understanding Image Classification Decisions and Improved Neural Network Robustness," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 508-516, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[32] Tao Zhuo et al., "Explainable Video Action Reasoning via Prior Knowledge and State Transitions," *Proceedings of the ACM International Conference on Multimedia*, pp. 521-529, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[33] K. Vanitha et al., "Enhanced Tuberculosis Detection using Vision Transformers and Explainable AI with a Grad-CAM Approach on Chest X-Rays," *BMC Medical Imaging*, vol. 25, no. 1, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Antonio Lavecchia et al., "Explainable Artificial Intelligence in drug Discovery: Bridging Predictive Power and Mechanistic Insight," *WIREs Computational Molecular Science*, vol. 15, no. 5, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[35] Mohammad Rasras et al., "Effects of Different Attention Mechanisms Applied on 3D Models in Video Classification," *International Conference on Computational Science and its Applications*, Istanbul, Türkiye, pp. 347-363, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[36] Pascal Benschop et al., "Motion Representations for Privacy-Aware Cross-Domain Action Recognition," *Frontiers in Imaging*, vol. 5, pp. 1-13, 2026. [CrossRef] [Google Scholar] [Publisher Link]