Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.066415

ARTICLE

ARNet: Integrating Spatial and Temporal Deep Learning for Robust Action Recognition in Videos

Hussain Dawood[1], Marriam Nawaz[2], Tahira Nazir[3], Ali Javed[2], Abdul Khader Jilani Saudagar[4,*] and Hatoon S. AlSagri[4]

[1]School of Computing, Skyline University College, Sharjah, 1797, United Arab Emirates
[2]Department of Software Engineering, University of Engineering and Technology-Taxila, Punjab, 47050, Pakistan
[3]Department of Software Engineering and Computer Science, Riphah International University-Gulberg Green Campus, Islamabad, 46000, Pakistan
[4]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11432, Saudi Arabia

*Corresponding Author: Abdul Khader Jilani Saudagar. Email: aksaudagar@imamu.edu.sa

Received: 08 April 2025; Accepted: 16 June 2025; Published: 31 July 2025

ABSTRACT: Reliable human action recognition (HAR) in video sequences is critical for a wide range of applications, such as security surveillance, healthcare monitoring, and human-computer interaction. Several automated systems have been designed for this purpose; however, existing methods often struggle to effectively integrate spatial and temporal information from input samples such as 2-stream networks or 3D convolutional neural networks (CNNs), which limits their accuracy in discriminating numerous human actions. Therefore, this study introduces a novel deep-learning framework called the ARNet, designed for robust HAR. ARNet consists of two main modules, namely, a refined InceptionResNet-V2-based CNN and a Bi-LSTM (Long Short-Term Memory) network. The refined InceptionResNet-V2 employs a parametric rectified linear unit (PReLU) activation strategy within convolutional layers to enhance spatial feature extraction from individual video frames. The inclusion of the PReLU method improves the spatial information-capturing ability of the approach as it uses learnable parameters to adaptively control the slope of the negative part of the activation function, allowing richer gradient flow during backpropagation and resulting in robust information capturing and stable model training. These spatial features holding essential pixel characteristics are then processed by the Bi-LSTM module for temporal analysis, which assists the ARNet in understanding the dynamic behavior of actions over time. The ARNet integrates three additional dense layers after the Bi-LSTM module to ensure a comprehensive computation of both spatial and temporal patterns and further boost the feature representation. The experimental validation of the model is conducted on 3 benchmark datasets named HMDB51, KTH, and UCF Sports and reports accuracies of 93.82%, 99%, and 99.16%, respectively. The Precision results of HMDB51, KTH, and UCF Sports datasets are 97.41%, 99.54%, and 99.01%; the Recall values are 98.87%, 98.60%, 99.08%, and the F1-Score is 98.13%, 99.07%, 99.04%, respectively. These results highlight the robustness of the ARNet approach and its potential as a versatile tool for accurate HAR across various real-world applications.

KEYWORDS: Action recognition; Bi-LSTM; computer vision; deep learning; InceptionResNet-V2; PReLU

1 Introduction

In computer vision (CV), HAR is employed to identify and classify human activities in videos or real-time camera visuals [1–3]. It is the process of automatically recognizing various activities, such as walking, running, jumping, and waving, from input samples [4]. The process of HAR includes various phases, such

as detecting and tracking individuals in the video, extracting relevant features from their movements, and classifying these features using various machine learning (ML) or deep learning (DL) approaches [5,6]. There exist various real-world applications of HAR across various domains, i.e., surveillance and security, where such systems can detect suspicious or anomalous behavior in public spaces, which can ultimately enhance communal safety and aid regulation implementation efforts [7]. Further, in sports analytics, such HAR systems can deliver a detailed analysis of the movements and actions of sportsmen, which can be used by the players for performance and strategy development. In addition, healthcare units can use such systems to look after elderly patients or those with movement disorders by recognizing specific actions that can generate alerts for caregivers to signify falls or other critical situations [8]. In the domain of human-computer interaction, HAR applications can provide more spontaneous interfaces that permit users to control devices through gestures and can enhance accessibility and user experience. The importance of such systems lies in their potential to reliably understand human behavior through visual data [9]. Accurately recognizing actions, such as HAR applications, can deliver valuable insights and real-time responses in numerous domains. In security, effective HAR systems deliver quicker and more accurate threat detection. However, in the areas of sports and healthcare, such systems translate to better performance monitoring and patient care, respectively [10]. In addition, due to the modern advances in technology, the integration of HAR systems with other applications can lead to the development of smart environments and innovative applications, such as driving progress in fields such as robotics, virtual reality, and others [11]. So, automated design HAR systems hold significant promise for improving safety, efficiency, and quality of life.

Researchers have employed a variety of advanced techniques for effective HAR, utilizing both traditional methods (ML) [12,13] and cutting-edge DL approaches [14,15]. Traditional ML approaches are based on hard-coded feature descriptors, such as Histogram of Oriented Gradients [16], Scale-Invariant Feature Transform (SIFT) [17], and Spatio-Temporal Interest Points [18]. Such computed features are later recognized into various groups with the help of numerous ML predictors such as Support Vector Machines (SVMs) [19] or Random Forests (RFs), and others [20]. These conventional approaches are very efficient in accomplishing CV tasks; however, they have significant limitations. Handcrafted features require extensive domain knowledge and manual tuning, which makes them less flexible and scalable. In the area of HAR, which comprises long video sequences, these approaches often fail to capture complex motion patterns and are sensitive to variations in lighting, viewpoint, and occlusion. Therefore, these methods show suboptimal HAR performance in real-world scenarios [21]. The field of HAR has been revolutionized with the advent of DL frameworks such as CNNs and Recurrent Neural Networks (RNNs) to overcome these limitations [1]. CNNs are utilized to automatically learn spatial features from video frames by eliminating the need for manual feature extraction [22]. These approaches are robust in learning complicated patterns and are effective in tackling the variations in the input data. RNNs, with their variants such as Long Short-Term Memory (LSTM) networks, are skillful at modeling temporal dependencies between frames and permit the effective capture of motion dynamics with time [23]. In addition, the latest approaches involve 3D CNNs, which compute spatiotemporal aspects of videos simultaneously by applying 3D convolutions over the video data. Such techniques permit the direct modeling of temporal progression within the spatial context of the video sequences [24]. Further techniques include Transformer approaches, which were originally designed for natural language processing and have been explored for HAR as well due to their capability to handle long-range dependencies and parallelize training [25]. Techniques such as Two-Stream Networks combine spatial and temporal information by processing RGB frames and optical flow separately before merging their representations and can effectively capture both appearance and motion cues. In literature, attention mechanisms have also been incorporated by researchers to focus on the most informative parts of a video sequence to enhance the ability of approaches to differentiate between similar actions and improve overall

accuracy. These diverse DL models are often combined and fine-tuned to drive the current advancements in HAR [24]. These methods deliver a more robust and precise understanding of human actions in diverse domains by significantly outperforming traditional methods and paving the way for more sophisticated and scalable solutions [14].

Despite significant progress in the area of DL for effective HAR applications, there remain several limitations, including one major issue is the requirement for large-scale labeled datasets to train such models effectively [26]. In addition, DL approaches such as those based on CNNs and RNNs have shown impressive performance; however, they are computationally intensive and require considerable hardware resources for both model tuning and inference. Such dense frameworks are a barrier for real-time applications or deployment on edge devices with limited processing power. Another limitation is the generalization capability of these models. DL models are sensitive to changes in the environment, such as variations in lighting, background, and viewpoint. Such techniques potentially lead to decreased performance when applied to different datasets or real-world scenarios separate from the training conditions [27]. These models struggle with occlusions and complex interactions involving multiple individuals and are unable to effectively learn delicate and complex dynamic interactions. Interpretability of the DL approaches is also another concern due to their black-box nature, making it difficult to understand their decision-making processes and diagnose errors. Lastly, the imbalanced data samples where specific actions are over-represented while others are under-represented can introduce model biases, which can affect their ability to recognize less frequent actions precisely. Therefore, addressing these limitations is crucial to ensure the reliable and robust delivery of the HAR systems in diverse real-world settings [7].

This study proposes an effective DL approach called ARNet to address the existing issues in this field. The ARNet framework utilizes both the spatial and temporal information of input samples and comprises two main modules, which are CNN and Bi-LSTM networks. The CNN part proposed a refined InceptionResNet-V2 architecture utilizing the parametric rectified linear unit (PReLU) activation approach in the convolution layers to effectively extract spatial features from individual video frames as the PReLU method uses learnable parameters to adaptively control the slope of the negative part of the activation function, allowing richer gradient flow during backpropagation for more robust information capturing and stable model training. These spatial features are then passed to the Bi-LSTM model to execute temporal analysis and understand the dynamic behavior of actions over time. Next, the ARNet model incorporates three additional dense layers to refine and nominate the relevant features. This step ensures a comprehensive learning of both spatial and temporal information. Finally, the computed features are passed to the classification part to execute the HAR task. This architecture of the ARNet assists it in accurately recognizing complex actions in videos and makes it a powerful tool for various HAR applications. The significant contributions of this study are listed below:

- The indicated ARNet architecture introduces a refined InceptionResNet-V2 architecture by incorporating the PReLU activation function in the convolutional layers. This innovation enhances the model's ability to learn a more detailed set of spatial features from videos at the frame level by addressing the limitations of the traditional activation function.
- The advantages of CNN for spatial feature extraction are combined with Bi-LSTM networks to embed spatiotemporal learning behavior in the ARNet architecture, which leads to a more comprehensive understanding of the dynamic behavior of actions over time.
- The incorporation of the additional dense layers in the ARNet model enhances the representation of extracted features and ensures a thorough capture of both spatial and temporal aspects, which ultimately enhances the capability of the model to recognize complex actions in videos.

- A strong experimental analysis is performed to confirm the robustness of the ARNet in effectively addressing existing issues in the field and confirm its scalability, recall, and generalization across various video datasets and transformation conditions.

The remaining sections are organized in the paper as follows: Section 2 discusses existing studies on AR, Section 3 provides details about the proposed ARNet, Section 4 introduces the datasets, parameters, and results, and Section 5 presents the conclusions. In addition, Table 1 provides the complete forms of the abbreviations used in this study.

Table 1: Abbreviations and their definitions

Abbreviation	Full form
CNN	Convolutional Neural Network
PReLU	Parametric Rectified Linear Unit
HAR	Human Action Recognition
ML	Machine Learning
DL	Deep Learning
CV	Computer Vision
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machines
LSTM	Long Short-Term Memory
RF	Random Forest
SSD	Single-shot detector
MLP	Multi-layer perceptron
MoCap	Motion capture
IMU	Inertial measurement units
NB	Naïve Bayes
DNN	Deep neural network
STO	Spider Monkey Optimization
STA	Spatio-temporal attention
V	Video sequence
f	Frames
y	Predicted value
s	Refined InceptionResNet-V2 values
z	Enhanced features
W	Weights
b	Biases
FC	Fully Connected
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
CM	Confusion Matrix
ELU	Exponential linear unit
SELU	Scaled exponential linear unit
LRCN	Long-term recurrent convolutional network

(Continued)

Table 1 (continued)

Abbreviation	Full form
CAM	Channel attention mechanisms
GRU	Gated recurrent unit
STHARNET	Spatio-temporal human action recognition network
CBVR	content-based video retrieval
AGOP	Adaptive Genetic Optimization Procedure

2 Related Work

This section analyzes the existing works performed for action recognition from visual data. The existing studies used for the said problem are broadly distributed into 2 types, termed conventional ML approaches and DL works.

Initially, ML approaches were discussed and used recently by the research community for performing HAR. Garcia-Gonzalez et al. [28] developed a conventional ML approach for HAR. For this, the work initially developed a dataset containing various actions performed by humans in the world. After this, various ML techniques were applied to the collected samples using different hyperparameter descriptions. The approach attains the best results using the random forest (RF) classifier with a maximum accuracy score of 92.97%; however, the results need improvement. Kapoor et al. [29] discussed another ML approach, where the researchers employed the OpenPose tool to compute the features of humans present in the visual sample. Next, for classification, the approach used various ML and DL classifiers such as SVM, LSTM, multi-layer perceptron (MLP), and others. The approach has attained an accuracy result of 87.77% along with the SVM predictor; however, the scores need enhancements. Azmat et al. [30] proposed an approach to perform the classification of various human activities. For this, the indicated approach first distributed the video into frames and used a bilateral filter to boost the area of interest. Next, the work employed quick shift segmentation to separate the human shape. Next, 13 skeleton features were computed along with location, angular relationships, and 3D point clouds. An expectation-maximization technique with a Gaussian mixture approach designed for elliptical groups. The focused points on the ellipses were traced throughout the activity. The NB optimizer was utilized to optimize the computed feature set, while a deep classifier was applied to categorize activities. The approach performs well for HAR; however, the approach needs to be evaluated on a more complex data sample to prove its robustness. Even though ML approaches have been extensively employed for HAR, these methods fail to perform well in real-world scenarios.

The robustness and effectiveness of the DL approaches have led researchers to utilize them for HAR. Cob-Parro et al. [31] projected a DL network for recognizing actions from the visual data. For this, the work employed a MobileNetv2-based single-shot detector (SSD) to locate people with various actions by drawing the bounding boxes around them. Next, the located actions are passed to the LSTM approach to perform the classification task. The method reported the highest precision of 99.28% on the KTH dataset; however, it was not effective in locating the smallest movements, such as teeth brushing. Zhang et al. [32] proposed a hierarchical video action classification model by designing a video-language learning approach. The model learns the relationships between various hierarchical video levels and employs a top-down constraint to enhance the accuracy of recognition predictions. The work performs well for action recognition; however, it is specifically designed for medical field-related actions and needs evaluations on a generic and standard data sample to prove its effectiveness. Disabling person activity recognition is vital for various clinical applications, including intensive care, epileptic seizure diagnosis, and home sleep monitoring. Precise

determination of in-bed actions from visuals is essential, but several challenges exist. These include the gap between lab and clinical settings, the need for non-intrusive nursing, and the restricted availability of labeled medical activity data. Therefore, Karácsony et al. [10] proposed an approach to focus on epileptic seizure classification and examine the challenges and trends in video-based in-bed monitoring, including monocular 3D motion capture (MoCap) and automated seizure classification. The work highlighted the potential of using 3D MoCap and skeleton-oriented HAR along with transfer learning to improve clinical diagnoses, though several issues remain, such as spatiotemporal permanency, tackling hidden objects, and robustness. Mekruksavanich et al. [33] proposed a DL model to address the challenges of recognizing activities when device positioning is uncontrolled. For this purpose, the approach proposed a progressive residual DL approach called Att-ResBiGRU, which performed well for location-dependent and independent HAR tasks. The approach is tested using 3 data samples and reported F1 scores of 86.69%, 96.23%, and 96.44% on the PAMAP2, REALWORLD16, and Opportunity datasets, respectively. The work performs well for HAR, with a huge computational burden. Khan et al. [34] indicated a DL model that utilized wearable devices, including inertial measurement units (IMUs), Ambient sensors, GPS, along Audio sensors, to accurately detect and classify human activities. Utilizing data from the Opportunity and Extrasensory data samples, the study introduced an advanced methodology with novel feature extraction techniques. The framework employed the GPS, audio, and IMU sensors to perform the localization task, while the IMU and Ambient sensors were applied for locomotion HAR. The methodology employed CNNs for recognizing indoor/outdoor actions and LSTM networks for locomotion activities. Evaluated with k-fold cross-validation, the system achieved classification results of 97% for locomotion on the Opportunity dataset, 89% on the Extrasensory dataset, and 96% for indoor/outdoor activities on the Extrasensory repository; however, it needs further performance improvements.

Khan et al. [8] discussed another DL model, where the dense keypoints are first calculated with the help of the CNN approach VGG19. Next, key points at varying angles were taken from horizontal and vertical gradients. Then, features from both previous steps were combined and later passed to nominate the relevant information based on relative entropy, mutual information, and strong correlation coefficient (SCC). At last, the selected optimal set of sample features was utilized to train the Naïve Bayes (NB) predictor to estimate the label. The approach was tested on various data samples and reported the highest accuracy value of 99.40% on the YouTube dataset; however, the work requires evaluation on a more challenging dataset. Kaya et al. [35] proposed a DL approach for HAR from the visual data. For this, the samples were initially preprocessed to make them appropriate for the DL approach. The processed data was then passed to the 1D-CNN approach to compute the visual information of samples and accomplish the classification of numerous human actions. The approach executes effectively for HAR; however, it is unable to tackle the unbalanced data. Kolkar et al. [36] developed a dense model for performing HAR from the video samples. For this, the approach is located around 156 temporal and frequency-oriented keypoints. Then, a deep neural network (DNN) approach with the Spider Monkey Optimization (DNN-SMO) strategy was proposed for categorizing human activities employing sensor data. The fitness method of the spider monkey was introduced in the hidden layer of the NN to boost classification results. Local and global leader fitness methods enhanced the information-capturing capabilities of the model, executing keypoint-level fusion and beforehand categorization. The work was evaluated using four different datasets, and the highest results were reported at 98.92% over the WISDM dataset; however, the results need further improvements. Brishtel et al. [37] also suggested a DL approach that transformed radar data into spectrograms for HAR. For this, the approach computed three maps, covering around 1 s, and processed them independently by employing the ResNet-18 approach. The keypoints computed from the last three frames were stored in a ring buffer, so only the newest frame needs processing when it arrives. The information captured from 3 frames was then combined and categorized

by a fully connected layer. The work also introduced a dataset for model evaluation; however, it needs evaluation on a standard sample to prove its robustness. Surek et al. [38] evaluated a DL approach employing the residual model with a vision transformer for performing HAR from the video samples. The approach shows an efficient solution for HAR; however, categorization performance needs improvement. Alhakbani et al. [39] employed a pre-trained CNN model, VGG16, for accomplishing HAR in an end-to-end way, where the utilized CNN approach computed a dense group of visual characteristics and performed the recognition task. The approach was evaluated using the KTH dataset, and an accuracy rate of 98% was reported. The approach performs well for HAR; however, it comes at the cost of increasing the computing burden. Many researchers have adopted the idea of using both spatial and temporal information of the videos for HAR, such as Dwivedi et al. [40] proposed a DL approach for HAR by using spatiotemporal sample analysis. Initially, Inceptionv3 was used for feature estimation from the input videos at the spatial level. The computed features were later communicated to the LSTM approach to perform the sequence analysis and accomplish the classification task. The approach reported the highest classification score of 98.87%, however, with a huge computing burden. The literature highlighted numerous DL works for unethical HAR, which effectively learn low-level time-based and pixel-level keypoints; however, face challenges with high-level visual information, limiting their performance. This issue leads to poor learning capabilities in deep learning models. In digital forensics, detailed video analysis is crucial for cybercrime examination and anticipation. For this, a spatiotemporal HAR approach was proposed in [41], which combined a 2-stream inflated 3D ConvNet (I3D) and spatiotemporal units. The I3D technique boosted 3D CNN performance by converting 2D conv windows into 3D, while the spatio-temporal attention (STA) unit enhanced knowledge by concentrating on the pixel and time-based information of each video frame. The approach was tested using four different datasets and attained the highest accuracy value of 97.20% over the NPDI dataset; however, the work needs extensive samples for training. A comparison of the existing approaches is provided in Table 2.

Table 2: Comparative analysis of the existing techniques

Reference	Technique	Dataset	Accuracy (%)	Limitations
[28]	Conventional features + RF	Custom dataset	92.97	Needs evaluation on a larger and standard dataset
[29]	OpenPose + SVM	Drone-Action dataset	87.77	The model needs evaluation on datasets with more complex human activities
[30]	13 skeleton features + Deep classifier	UAVGesture DroneAction UAVHuman dataset	95 90 44	The approach needs to enhance its generalization power
[31]	MobileNetv2-based SSD detector + LSTM	KTH	98	The model is not effective in locating the tiny movements
[32]	Hierarchical features	Custom dataset	98	The work is designed for medical field-related actions only
[33]	Att-ResBiGRU	PAMAP2 REALWORLD16	96.61 96.11	The work is computationally complex

(Continued)

Table 2 (continued)

Reference	Technique	Dataset	Accuracy (%)	Limitations
[34]	CNN + LSTM	Opportunity datasets	90.27	The work needs further performance improvements
		Opportunity	97	
		Extrasensory dataset	89	
[8]	SCC + NB	YouTube dataset	99.40	The approach needs evaluation on a more advanced dataset
[35]	1D-CNN	UCI-HAPT	96.90	The work is unable to tackle the unbalanced data
[36]	156 temporal and frequency-oriented features + DNN-SMO	WISDM	98.92	The work requires performance improvements
[37]	ResNet-18	Custom dataset	71.30	The work requires performance improvements and testing on a more challenging dataset
[38]	Vision transformer	HMDB51	41.90	The approach needs further enhancements
[39]	VGG16	KTH	98	The work is computationally complex
[40]	Inceptionv3 + LSTM	Custom dataset	98.87	The work lacks capturing long-range dependencies in visual data
[41]	I3D	NPDI	97.20	The work needs extensive samples for training

The comparison provided in Table 2 indicates that while various ML, DL, and hybrid approaches achieve competitive accuracy, several challenges persist in the domain of HAR. Common limitations include reliance on handcrafted or shallow features, poor generalization to complex or diverse datasets, computational inefficiencies, inability to model long-range dependencies, and limited integration of spatial and temporal information. In addition, some high-performing models require extensive training data or are designed for specific application domains, restricting their scalability and adaptability. The proposed ARNet framework introduces a unified and robust approach that combines the strengths of both CNNs and RNNs for comprehensive spatial-temporal modeling to address these limitations. First, the spatial feature extractor is built on a refined InceptionResNet-V2 architecture that integrates PReLU activation functions within convolutional layers, enabling adaptive learning of complex spatial features while maintaining model stability and efficient training. This directly mitigates the issues of shallow feature extraction and poor adaptability to visual variations such as lighting, background clutter, and subtle motions. Second, the Bi-LSTM module captures long-range temporal dependencies in both forward and backward directions, enabling the model to understand dynamic action sequences more effectively than models relying solely on short-term temporal features. Third, to enhance representation and decision-making capacity, three dense

layers are appended after the Bi-LSTM, supporting deeper fusion and refinement of learned features. This layered integration not only ensures effective spatial-temporal cooperation but also raises scalability across varied datasets. In addition, ARNet is trained and evaluated with cross-corpus testing to demonstrate its generalization power, making it suitable for real-world applications such as surveillance, healthcare, and human-computer interaction.

3 Proposed Method

This study introduces a novel framework called the ARNet architecture, which is designed for robust HAR in video sequences. The approach comprises two main modules, which are the CNN and a Bi-LSTM network. The CNN part proposed a refined InceptionResNet-V2 architecture by utilizing the PReLU activation approach in the convolution layers to effectively extract spatial features from individual video frames. These spatial features are then passed to the Bi-LSTM model to execute temporal analysis to understand the dynamic behavior of actions over time. Next, the ARNet model incorporates three additional dense layers to refine and nominate the relevant features. This step ensures a comprehensive learning of both spatial and temporal information. Finally, the computed features are passed to the classification part to execute the HAR task. Such architecture of the ARNet assists it in accurately recognizing complex actions in videos and makes it a powerful tool for various HAR applications. A detailed overview of the ARNet is provided in Fig. 1, and detailed steps are explained in Algorithm 1.

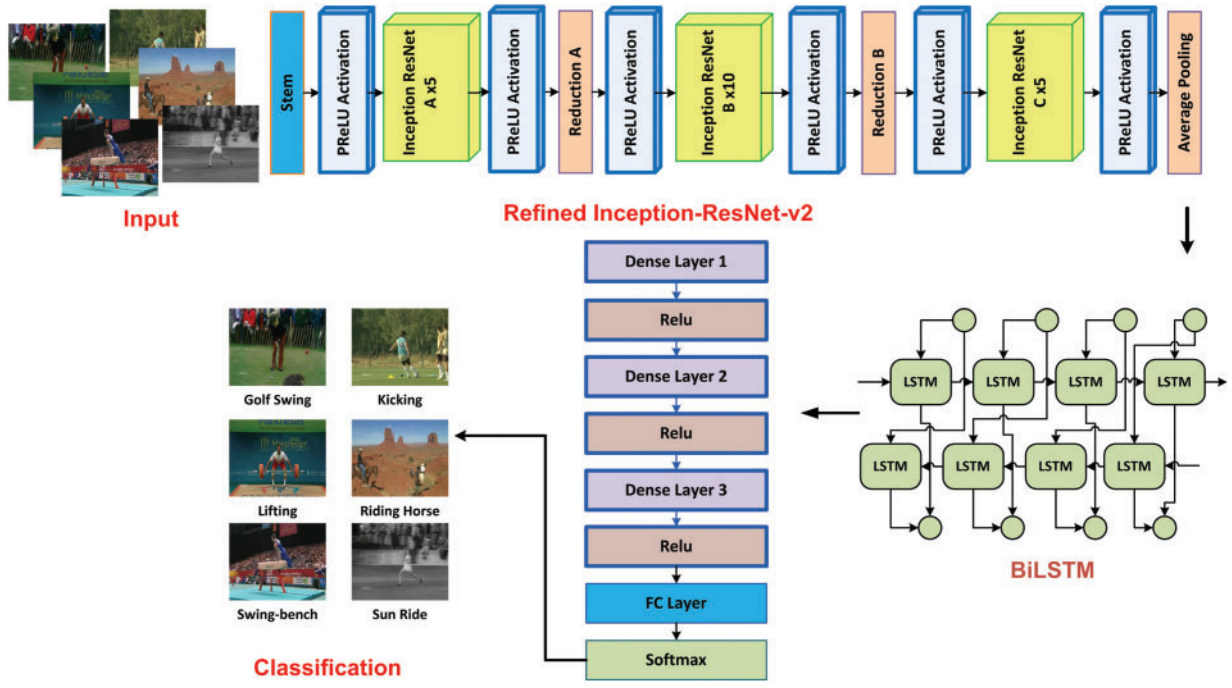

Figure 1: The workflow of the ARNet

Algorithm 1: ARNet—human activity recognition framework

Input: A video sequence $V = \{f_1, f_2, ..., f_n\}$ where f_i are video frames
Output: Predicted activity label y
//Preprocessing:
For each frame $f_i \in V$:
Resize and normalize f_i.
//Spatial Feature Extraction:
For each frame $f_i \in V$: //f = *frames*, V = *video sequence*
$s_i \leftarrow$ Refined InceptionResNet-V2(f_i) using PReLU activation //s = *refined inceptionresnetv2 values*
//Temporal Feature Learning:
$T \leftarrow$ Bi-LSTM($\{s_1, s_2, ..., s_n\}$) //T = *temporal features*
//Feature Enhancement via Dense Layers:
$z_1 \leftarrow \text{ReLU}(W_1 T + b_1)$ //z_1 = *enhanced features of dense layer-1*
$z_2 \leftarrow \text{ReLU}(W_2 z_1 + b_2)$ //z_2 = *enhanced features of dense layer-2*
$z_3 \leftarrow \text{ReLU}(W_3 z_2 + b_3)$ //z_3 = *enhanced features of dense layer-3*
//Classification:
$y \leftarrow \text{Softmax}(W_f z_3 + b_f)$ //y = *predicted values*
Return the predicted label y

3.1 CNN Module

The work focuses on proposing an automated framework for HAR in video sequences called the ARNet approach. The first part of the ARNet is a CNN model that utilizes a pre-trained InceptionResNet-V2 [42] model. The utilization of the pre-trained CNN approach is advantageous in several ways, as it utilizes the previously learned knowledge from a large data sample such as ImageNet to solve a new problem such as HAR. The pre-trained CNNs are skillful at learning hierarchical visual aspects by initially computing low-level features such as edges and textures, then developing to middle-level features such as shapes and patterns and eventually taking high-level features that represent complex objects and scenes. Therefore, using the capabilities of the InceptionResNet-V2 model in the ARNet architecture ensures that the proposed model can effectively capture the spatial patterns present in each video frame. These feature capabilities are mandatory for the reliable implementation of HAR, as low-level features help identify basic motion and contours, middle-level features assist in recognizing specific body parts and their movements, and high-level features enable the understanding of complex interactions and activities. So, the employment of a pre-trained model in the ARNet architecture enables it to vigorously process and interpret the complicated details of each frame and ensure robust spatial feature extraction. A view of this process is presented in Fig. 2.

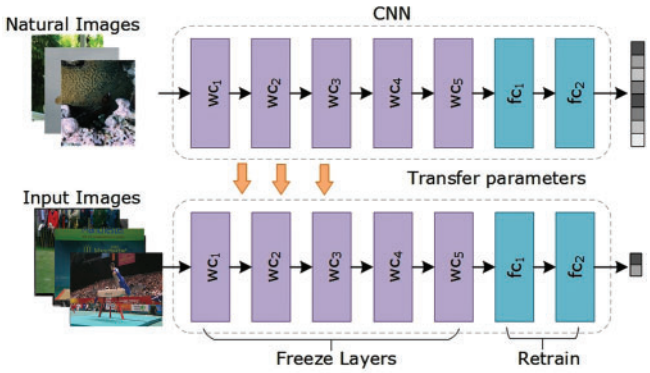

Figure 2: A view of the transfer learning procedure

3.2 Refined Inception-ResNet-V2

The proposed ARNet model includes a refined Inception-ResNet-V2 CNN framework by replacing the ReLU activation approach with PReLU in the convolution layers of the base models. The main reason for selecting the Inception-ResNet-V2 as the spatial-level information-capturing module of the ARNet is its impressive results in extracting hierarchical features from visuals, which is demanding for complicated tasks such as HAR. The InceptionResNet-V2 framework joins the strengths of Inception and residual connections and presents a powerful combination that boosts feature extraction and network training efficiency. The employed CNN architecture is well-suited to the problem (HAR) as it outperforms in capturing both fine-grained facts and sophisticated semantic features of long video sequences. The InceptionResNet-V2 approach is composed of 3 major blocks called the Inception-ResNet-A, Inception-ResNet-B, and Inception-ResNet-C units, respectively (Fig. 3). These blocks perform an important role in capturing the relevant information for HAR, such as the Inception-ResNet-A block, which mainly focuses on capturing local patterns within the input frames. This unit comprises several parallel convolutional layers with variable window sizes, which enables it to simultaneously compute different pixel frequencies. The network effectively learns a dense set of low-to-mid-level keypoints by combining these features through concatenation. The ability of A-block to capture the frame information at diverse scales empowers it to recall complex visual movements and local changes, which are critical for reliable HAR applications. Next, the B-block takes the characteristics extracted by the first block and focuses on learning more abstract, mid-level features. The B-block of the InceptionResNet-V2 model incorporates residual links, which assist the model in mitigating the vanishing gradient problem and facilitate the training of deeper networks. The residual links permit the model to acquire identity mappings and capture complex patterns and relationships between features. Such behavior of the B-block of the model helps the network to understand the temporal context and relationships between consecutive frames, which ultimately boosts the ARNet's power to recognize dynamic actions. Last, the C-block of the model captures high-level semantic information by using broader convolutional layers and residual links to extract complex representations of the input frames. This unit learns high-level patterns and interactions between different objects and classes within the input samples. Further, the refined Inception-ResNetv2 architecture allows the ARNet to take advantage of a comprehensive feature computation process that spans multiple levels of abstraction. The combined strengths of inception modules and residual connections enable ARNet to efficiently and effectively capture the spatial nuances of each frame. Further, the employment of the PReLU activation boosts the ability of the CNN module to extract a more powerful set of sample features, which is described in detail in the later section. A representation of the refined Inception-ResNet-V2 is given in Fig. 4.

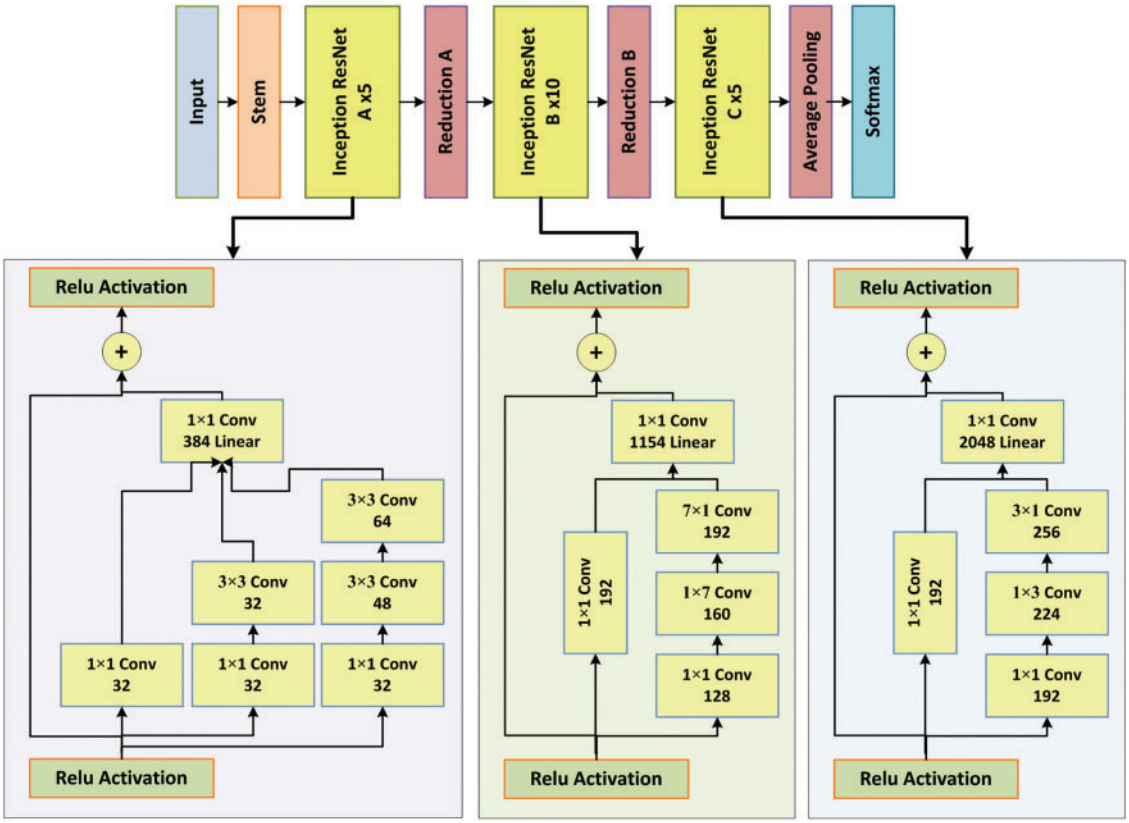

Figure 3: Refined Inception-ResNet-V2 approach

3.2.1 Convolution Layer

The convolution layers of the refined Inception-ResNet-V2 approach are focused on extracting the dense spatial information from a given visual sample, which is numerically explained in Eq. (1):

$$V_u^T = f\left(\sum_{v \in M_i} \left(S_{vu}^T * V_v^{T-1} + \beta_u^T\right)\right) \tag{1}$$

where T is the total layers of the model, and V and $*$ are the keypoints vector with window S, and the convolution operation. Further, β is the biased component, and M_i is the keypoint maps. As per network requirements, the size of all frames is set to 229 × 229.

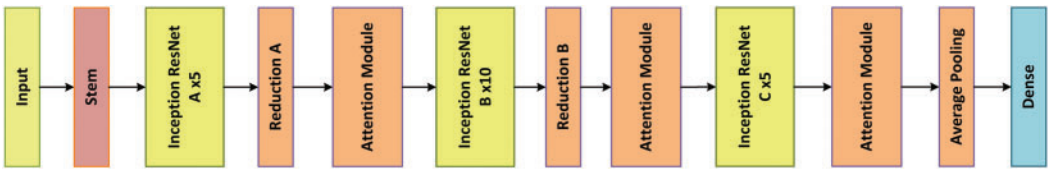

Figure 4: (Continued)

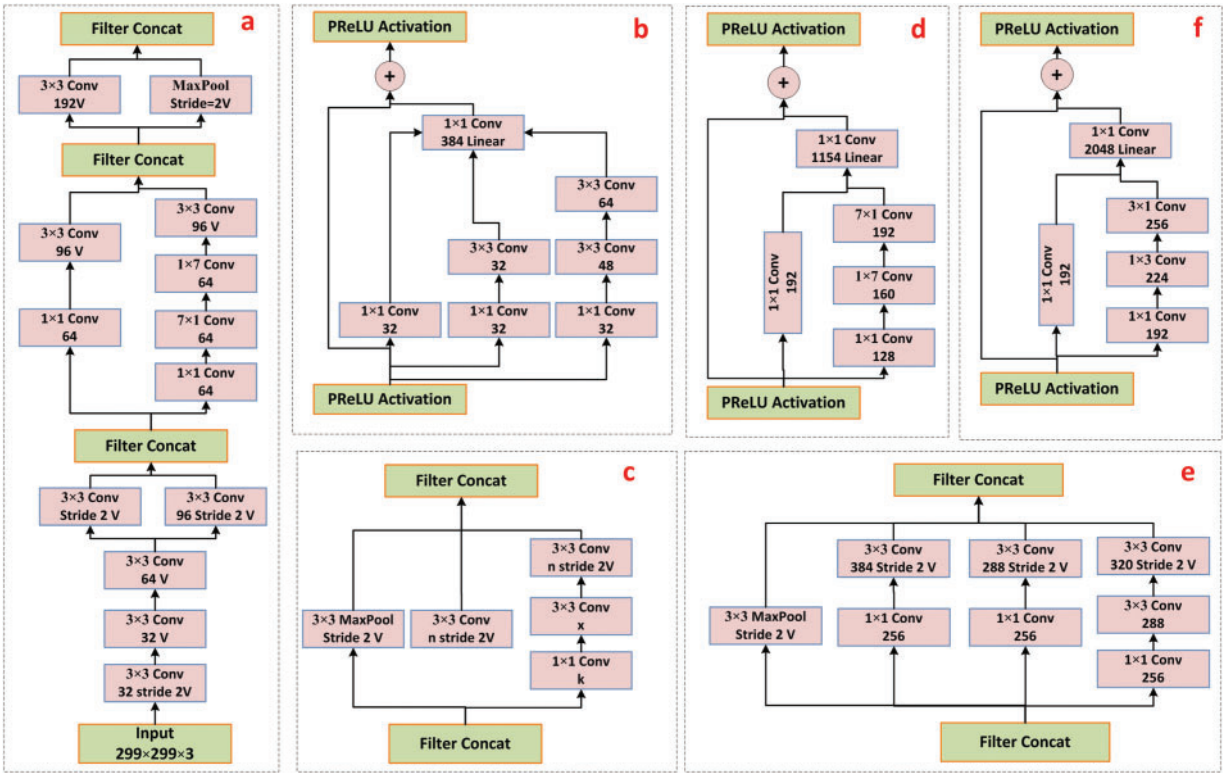

Figure 4: A thorough representation of the refined Inception-ResNet-V2 model: (**a**) stem block, (**b**) A block, (**c**) reduction unit, (**d**) B block, (**e**) reduction B, and (**f**) C block

3.2.2 Activation

The implementation of ARNet is selected to enhance the feature extraction capabilities of the Inception-ResNetv2 architecture by replacing the ReLU activation function with the PReLU in the convolutional layers. This decision was motivated by PReLU's ability to learn adaptive parameters for each neuron during training, which can help mitigate the issue of dying neurons commonly associated with ReLU. In comparison to ReLU, which sets negative scores to zero, the PReLU method employs learnable parameters that adaptively regulate the slope of the negative part of the activation function. The capability of the PReLU method to hold minute negative scores permits it to hold a richer gradient flow in the phase of backpropagation. Such behavior of this activation method causes more effective and stable training of DNNs. Further, it boosts the model's empowerment in computing a diverse and complex set of sample information. Such behavior of the activation method is beneficial for HAR, where taking refined disparities and patterns in video frames is vital. Therefore, the PReLU function assists the ARNet approach in better discriminating between different actions and enhancing the overall discriminative power of the model by holding and seeking information from the negative scores. In addition, PReLU can improve model generalization by reducing overfitting and allowing the network to learn more robust features from the data. Such characteristics are important for HAR applications where the designed automated systems must accurately classify diverse and potentially complex actions across different environments and conditions. So, the introduction of the PReLU into the activation layers of the ARNet model boosts its ability to extract meaningful visual patterns and leads to more accurate and reliable action recognition performance.

$$PReLU(i) = \max(0, i) + a \times \min(0, i) \tag{2}$$

where i is the input to the activation function, and a denotes the learnable parameter that controls the slope of the negative part of the function. The design guarantees that PReLU holds the positive activations unchanged by permitting negative activations to be scaled by the parameter a, by introducing flexibility and compliance in the activation behavior. A comparative analysis of the graphs of ReLU and PReLU activation functions is shown in Fig. 5.

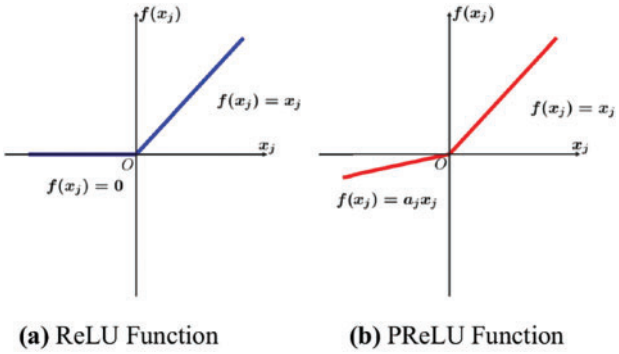

Figure 5: Comparison of ReLU vs. PReLU [43]

3.2.3 Pooling Layer

In the next phase, right after the convolution layers, an average pooling layer is introduced in the ARNet, which aggregates spatial information along every feature map and assists in reducing the dimensionality by keeping key features. For a 2D input feature map X, the average pooled output Y at position (i, j) with a window size $k \times k$ is computed as:

$$Y(i,j) = \frac{1}{k^2} \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X(i+m, j+n) \tag{3}$$

where (i, j) is the top-left coordinate of the pooling window in the input feature map. The final output of this part is a feature vector with dimensions of 1536 per frame, holding important pixel-level information that is later fed into the Bi-LSTM unit for temporal examination.

3.3 Bi-LSTM

After the pooling layers, the next layer is the Bi-LSTM module in the ARNet architecture [44]. The Bi-LSTM module inherits the competencies of RNNs and focuses on processing sequential information by preserving a hidden state that changes over time. The conventional RNNs are open to the vanishing gradient problem, where gradients diminish exponentially over long sequences. Therefore, RNNs are not proficient in handling long-term dependencies, such as in the case of HAR. Therefore, this study utilized the bi-directional LSTM approach in the ARNet, which is an extended form of the LSTM and was originally designed to overcome the problems. Initially, the LSTM approach tackles the RNN's issues by including gating mechanisms, i.e., input, forget, and output gates that control the information flow in the system. Such an architectural description of the LSTM allows it to remember or forget nominated past information and makes it appropriate for jobs demanding the modeling of time-based dynamics over prolonged sequences.

The final keypoints vector computed by the CNN unit is described as k_t, and the hidden state and the memory cell are denoted by C_{t-1} and b_{t-1}. The numeric implementation of LSTM is provided in Eqs. (4)–(8).

$$p_t = \sigma(\omega_{kp}k_t + \omega_{cp}C_{t-1} + \omega_{bp}b_{t-1} + B_p) \tag{4}$$

$$g_t = \sigma(\omega_{kg}k_t + \omega_{cg}C_{t-1} + \omega_{bg}b_{t-1} + B_g) \tag{5}$$

$$b_t = g_t b_{t-1} + p_t tanh(\omega_{kb}k_t + \omega_{cb}C_{t-1} + \beta_b) \tag{6}$$

$$o_t = \sigma(\omega_{ko}k_t + \omega_{co}C_{t-1} + \omega_{bo}b_t + B_o) \tag{7}$$

$$C_t = o_t.\tanh(b_t) \tag{8}$$

where σ is the sigmoid activation function, t to time, and p, g, o, and c are the input, forget, output gates, and memory cell states, respectively. Further, the ω corresponds to weights and B to biases.

However, LSTMs have limitations in capturing complex temporal relationships in action recognition tasks. They struggle with modeling very long-term dependencies and are prone to forgetting earlier context when processing lengthy video sequences. In addition, LSTMs function consecutively and lack parallel processing abilities, which causes slower model tuning and inference times for comprehensive video datasets. The Bi-LSTM module of the ARNet model overcomes these challenges by processing video samples bidirectionally and simultaneously including visual information from both past and future settings. Such a model architectural description allows it to compute complex progressive attributes inherent in action sequences and boost the robustness of HAR. A view of the Bi-LSTM module is provided in Fig. 6.

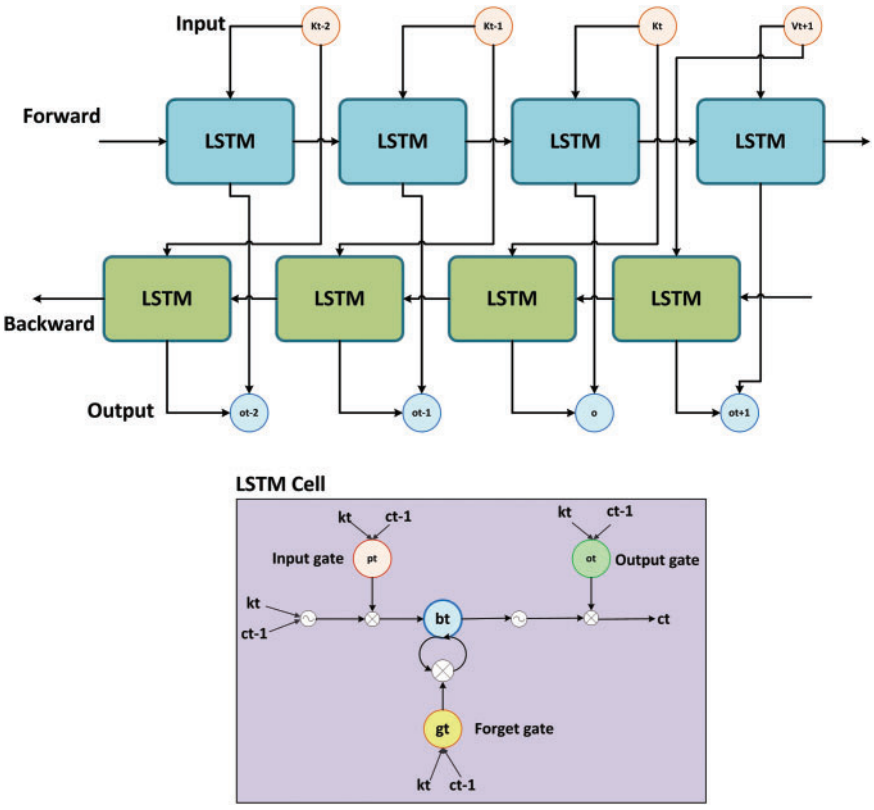

Figure 6: A pictorial view of the structure of the Bi-LSTM model

This part contains two categories of hidden units called the forward (C_t^r) and backward (C_t^w) states. The C_t^r analyzes video dynamics in a forward manner of time, i.e., $t = 1, 2, 3, \ldots, T$, and the C_t^w in a backward manner as $t = T, T-1, \ldots, 1$. Finally, the outcome o_t is calculated by joining the values computed from C_t^r, and C_t^r. The mathematical elaboration of Bi-LSTM is provided in Eqs. (9)–(11).

$$C_t^r = \tanh(\omega_{kc}^r k_t + \omega_{cc}^r C_{t-1}^r + B_c^r) \tag{9}$$

$$C_t^w = \tanh(\omega_{kc}^w k_t + \omega_{cc}^w C_{t+1}^w + B_c^w) \tag{10}$$

$$o_t = \omega_{cc}^r C_t^r + \omega_{cc}^w C_t^w + B_o \tag{11}$$

3.4 Dense Layers

After the Bi-LSTM unit, the ARNet architecture includes three dense layers, along with the ReLU activation method and a dropout layer. The primary purpose of these added layers is to boost the empowerment of the approach and highlight visual characteristics relevant to HAR by eliminating noise and unwanted background data. This modification in the ARNet model enables it to accurately detect and classify actions, particularly with diverse transformation settings such as fluctuations in lighting, color, and pixel locations of key action points. So, by including these layers, the proposed model can effectively integrate and refine earlier computed visual information and improve its capacity to differentiate between diverse groups of human actions. Let $o \in Rn$ represent the output feature vector from the Bi-LSTM module. The dense layers operate as follows:

$$Z_1 = W_1 o + b_1, h1 = ReLU(Z_1) \tag{12}$$

$$Z_2 = W_2 h1 + b_2, h2 = ReLU(Z_2) \tag{13}$$

$$Z_3 = W_3 h2 + b_3, h3 = ReLU(Z_3) \tag{14}$$

where W and b are the weight matrices and bias vectors for each dense layer. At last, a small dropout of 0.25 is added to alleviate overfitting issues by haphazardly disabling neurons in the phase of model tuning, which further assists the model in enhancing its robustness and generalization. After this, the computed information is passed to the last layer, which is described in the subsequent section.

3.5 Fully Connected (FC) Layer

The last part of the ARNet architecture is an FC layer containing a softmax activation designated to execute the action categorization task. The main task of this layer is to calculate the likelihood spread along predetermined action groups based on the information captured by the previous layers. The activation method, which is softmax in this case, ensures that the final probabilities summarize to 1, simplifying interpretation and decision-making in HAR. The numeric form of the softmax is given in Eq. (15).

$$\delta(O_x) = \frac{\exp(O_x)}{\sum_{q=0}^{n-1} \exp(O_m)} \tag{15}$$

where (O_x) and (O_q) are the final and input vectors, and q shows the corresponding number of classes.

3.6 Loss Method

ARNet's softmax layer employs the cross-entropy loss method [45] to measure the disparity between predicted action probabilities and the actual labels during training. Cross-entropy loss is particularly suited for multi-class classification tasks such as action recognition, where it penalizes incorrect classifications more

severely, encouraging the model to output confident probabilities for the correct action classes. Further, it is proficient in handling the class imbalance problem. Mathematically, the cross-entropy loss F is defined as:

$$F = \frac{1}{N}\sum_{j=1}^{q}\log\left(\frac{e^{s_q}}{\sum_k e^{s_j}}\right) \tag{16}$$

where q is neurons in the final layer, and s_q is the input vector.

4 Results

The datasets are utilized to test the model performance, the parameters employed for measuring results, and a detailed discussion of the obtained scores.

4.1 Dataset

For the tuning and testing of ARNet, this study utilized three widely recognized action recognition datasets: HMDB51 [46,47], UCF Sports [48,49], and KTH [50,51]. The HMDB51 dataset consists of 6766 video clips spanning 51 distinct action categories, including various everyday activities such as running, eating, and dancing. The samples for this dataset are collected from movies, online available clips, and various other sources, providing a diverse and complex set of human actions captured in a wide range of environmental conditions such as varying camera angles, lighting conditions, and backgrounds. The UCF Sports data sample contains a total of 150 visual sequences of sportspersons carrying out various sport-related actions such as diving, golf swinging, and pole vaulting. This dataset is considered due to the varying nature of sports actions present in this dataset, with cluttered backgrounds, which make it a challenging dataset in this domain. Finally, the KTH action repository contains 600 videos from 6 daily actions performed by humans in real-world scenarios such as walking, jogging, running, boxing, hand-waving, and handclapping. The visuals of this sample were taken under controlled settings with the same backgrounds and a fixed camera. This study performed a thorough evaluation of the ARNet approach to show its robustness against a wide spectrum of HAR by taking datasets of diverse nature, guaranteeing vigorous results and generalization competencies in real-world trends.

4.2 Performance Measurement Parameters

Several standard measures, such as accuracy and true positive rate (TPR), are computed to review the behavior of the model for HAR. For the problem under analysis, TPR indicates the ratio of the number of appropriately recognized actions to the total actions that belong to the group (both true positives and false negatives). TPR computed the capability of the approach to classify actions from a specific group. The mathematical formula for computing TPR is given in Eq. (17).

$$TPR = TP/(TP + FN) \tag{17}$$

where

- TP (True Positives) is the number of actions correctly identified as the true class.
- FN (False Negatives) is the number of actions that belong to a particular class but were incorrectly identified as a different class.

Accuracy in HAR is the ratio of the number of correctly identified actions (both true positives and true negatives) to the total number of actions. Accuracy is described in Eq. (18).

$$Accuracy = (TP + TN)/Total\ Actions \tag{18}$$

where

- *TP* (True Positives) is the number of actions correctly identified as the true class.
- *TN* (True Negatives) is the number of actions correctly identified as not belonging to a particular class.

Total Actions is the sum of all actions, including true positives, true negatives, false positives, and false negatives. This can be expanded to:

$$Total\ Actions = TP + TN + FP + FN \tag{19}$$

FP (False Positives) are the actions incorrectly identified as a particular class. FN (False Negatives) are actions that belong to a particular class but are incorrectly identified as being from a different class.

4.3 Model Evaluation

This section describes the recognition results attained by the ARNet approach for all employed datasets named HMDB5, KTH, and UCF Sport. Initially, the classification accuracy of the approach over all three datasets is discussed by plotting the results in Fig. 7. The results indicate that the proposed approach performs effectively for all three datasets. The high performance of ARNet across these diverse datasets highlights its robustness and adaptability. In the case of the HMDB51 data sample, which comprises a diverse set of human actions from 51 classes, the proposed approach, ARNet, achieves an average accuracy of 93.82% by successfully learning both the pixel and time-based information of samples. Further, for the KTH action repository, the proposed approach attains a classification score of 99%, showing its improved recognition ability to recall the action classes present in this dataset. Lastly, in the UCF Sports data sample, which contains diverse activities related to numerous sports, the model obtains a categorization score of 99.16%, which indicates the generalization power of the proposed approach in learning quite different types of HAR and demonstrates its applicability to real-world applications.

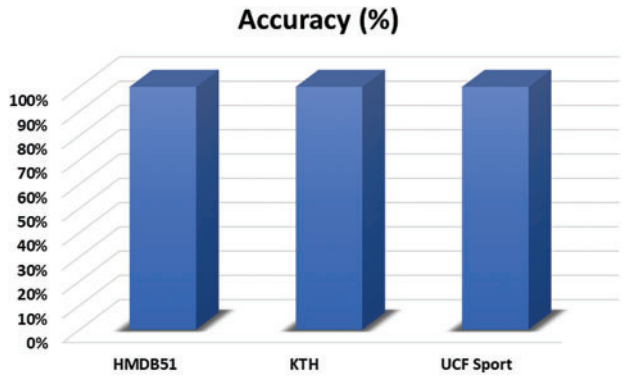

Figure 7: ARNet classification accuracy over all employed data samples

A detailed analysis of the evaluation metrics comprising Precision, Recall, and F1-Score is conducted across three benchmark datasets named KTH, UCF Sports, and HMDB51 to further validate the effectiveness and robustness of the proposed ARNet model. The results are shown in Table 3, which demonstrates the strong discriminative ability of ARNet in recognizing diverse human actions. On the KTH dataset, the model achieved a Precision of 99.54%, a Recall of 98.6%, and an F1-score of 99.07%, indicating its high accuracy and minimal false positives in a relatively controlled environment. For the UCF Sports dataset, ARNet maintained consistent performance with 99.01% Precision, 99.08% Recall, and a balanced F1-Score

of 99.04%, reflecting its capability to generalize well to real-world sports scenarios involving complex poses and backgrounds. On the more challenging HMDB51 dataset, which contains a wide range of actions and diverse video conditions, ARNet achieved 97.41% Precision, 98.87% Recall, and an impressive F1-Score of 98.13%, showing the effectiveness and robustness of the proposed approach in handling noisy and varied inputs. These strong performance metrics across datasets confirm ARNet's ability to precisely and reliably detect and classify human actions with both high confidence and comprehensive coverage.

Table 3: Performance comparison of the proposed work in terms of precision, recall, and F1-Score

Dataset	Precision (%)	Recall (%)	F1-Score (%)
KTH	99.54	98.6	99.07
UCF Sports	99.01	99.08	99.04
HMDB51	97.41	98.87	98.13

Now, the study details the results for all employed datasets, i.e., HMDB51, KTH, and UCF Sports, by discussing the confusion matrix (CM), which is a vital tool in the computer vision domain to explain the classification scores. The CM can provide a thorough analysis of results by showing the right and misclassified predictions, permitting group-wise accuracy evaluation. Such analysis enables the researchers to locate the classes that are often misclassified, allowing analysis of the approach's weak areas and indicating improvements. Fig. 8 depicts the CM obtained for the KTH action sample, which exhibits that the ARNet approach performs well for all six groups of this sample. The ARNet approach obtains an average TPR of 98.60%, which shows the robustness of the proposed approach.

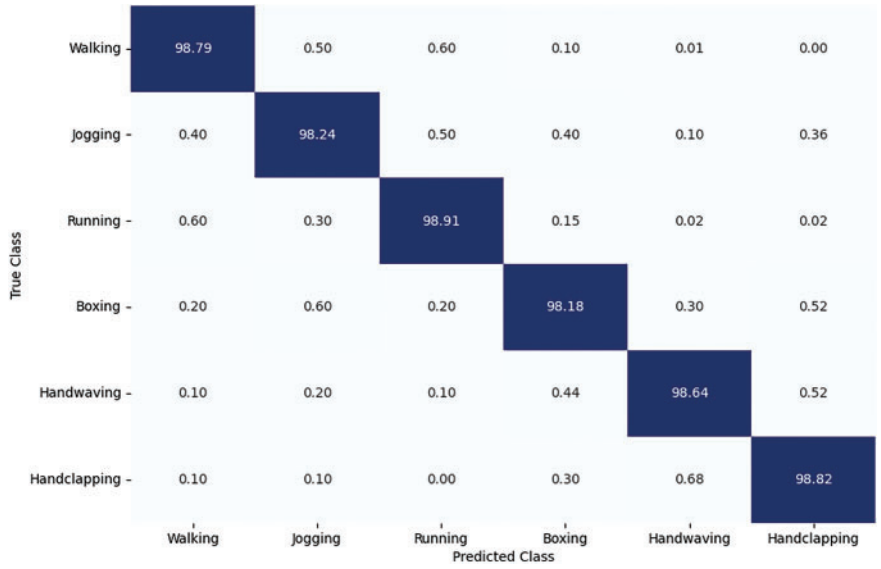

Figure 8: ARNet attained CM over the KTH dataset

Fig. 9 shows the CM obtained by ARNet for the UCF Sports dataset. It indicates that ARNet achieves remarkably good scores for all 11 categories of this dataset. Further, the scores given in Fig. 9 show that the ARNet approach attains the lowest TPR value of 98.09% for the kicking group, with the highest

misclassification rate of 0.7% in the kicking and Golf Swing groups. This is because the approach develops a minor confusion among these groups because of the resemblance in the dynamic lower-body arrangements of these actions. The ARNet reports the largest TPR for the skateboarding class, with a value of 99.45%. This highlights the robustness of the approach in recognizing dissimilar and complex actions related to skateboarding. Collectively, the proposed model achieves a TPR of 99.08%, underlining its effectiveness in precisely remembering all classes. The high recall value against diverse action classes confirms the ARNet model's ability to diminish false negatives and guarantee reliable HAR, making it an influential tool for practical areas in video analysis and surveillance.

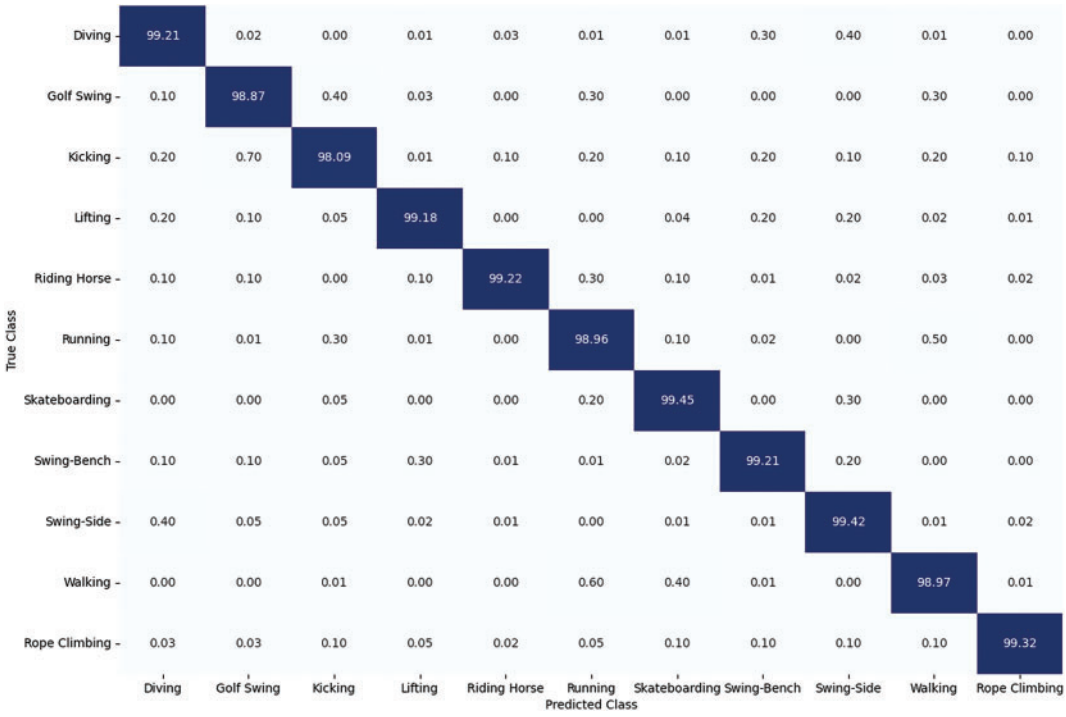

Figure 9: ARNet attained CM over the UCF Sports dataset

Fig. 10 illustrates the CM attained by the ARNet over the HMDB51 samples. The scores provided in Fig. 10 indicate the outstanding performance of the proposed approach in successfully remembering all 51 groups present in this large and challenging human action data sample. An average TPR is 98.87%, demonstrating its robustness for such a diverse set of human actions. The results provided in terms of CM against all three employed datasets indicate that the proposed approach can lessen false negatives, guaranteeing that most human actions are correctly identified. The results assure the generalization, high recognition, and effectiveness of the ARNet for HAR, marking it a powerful solution for various related tasks in real-world scenarios.

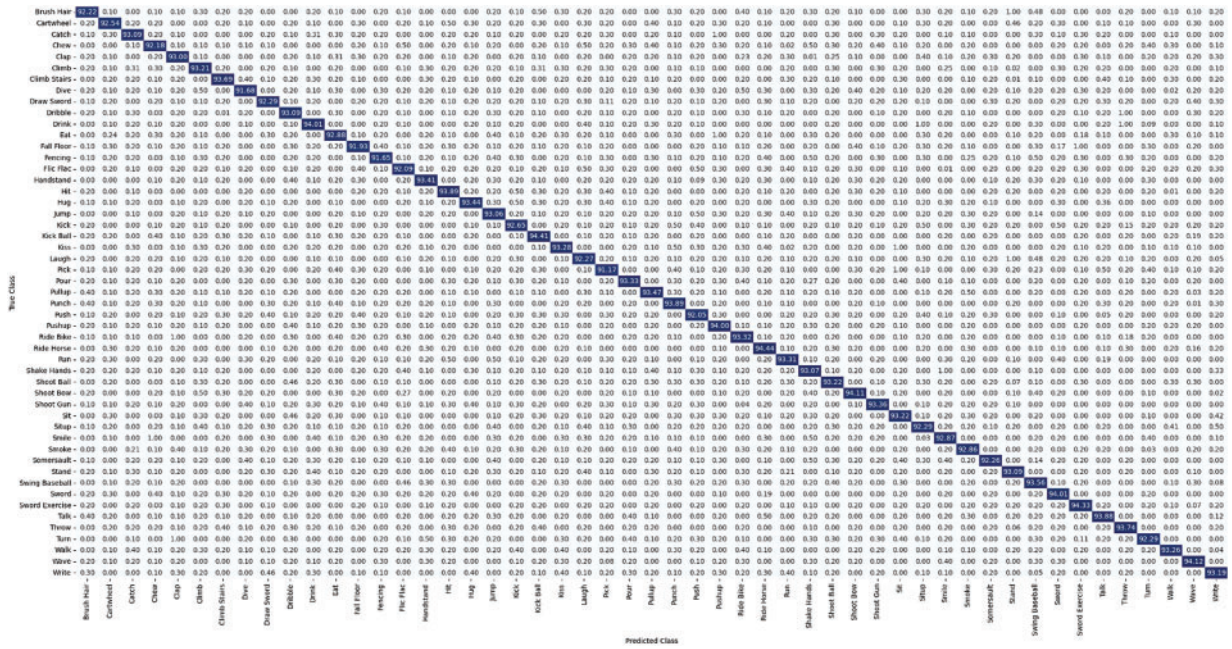

Figure 10: Confusion matrix attained by the ARNet over the UCF Sports dataset

4.4 Ablation Study

This section compares the performance of the ARNet against several activation approaches utilized among the convolution layers of the InceptionResNet-V2 approach. Hence, the performance of the proposed approach is evaluated with the original ReLU method, LeakyReLU, exponential linear unit (ELU), and scaled exponential linear unit (SELU) activation functions. The attained accuracy comparison for all three employed data samples is listed in Table 4. The results show that the proposed approach attains the highest result compared to the PreLU-based activation approach. The superior performance of PReLU over other activation functions is attributed to its ability to adaptively learn the parameters of the activation function, thus mitigating the issues of dying neurons common with ReLU. Unlike LeakyReLU, which uses a fixed slope for negative values, PreLU allows for the slope to be learned, providing greater flexibility and improved performance. ELU and SELU provide better performance than ReLU by addressing the vanishing gradient problem and providing self-normalization properties, respectively. However, PreLU's adaptive nature allows it to outperform ELU and SELU by dynamically adjusting to the specific characteristics of the data, leading to higher accuracy across all datasets.

Table 4: ARNet performance comparison to various activation methods

No.	Activation method	Accuracy (%) HMDB51 dataset	Accuracy (%) KTH dataset	Accuracy (%) UCF Sports dataset
1.	ReLU	89.94	92.10	93.23
2.	LeakyReLU	91.82	93.87	93.68
3.	ELU	90.24	94.07	94.09
4.	SELU	91.93	95.61	95.66
5.	**Proposed (PReLU)**	93.82	99.00	99.16

4.5 Comparison to DL Methods

The section compares ARNet with various well-known DL approaches for all three employed datasets. First, results are discussed for the HMDB51 data sample. Several DL models such as GoogleNet, Inception, as mentioned in [52], ResNet101 [53], 2stream-Inceptionv1 [54], and EfficientNetB3, Xception, as provided in [55] are considered, and the obtained comparison is given in Table 5. The values in Table 5 show that the proposed approach performs better than other DL methods and attains the highest recognition results. The superiority of ARNet is evident through its ability to effectively integrate spatial and temporal features, optimizing the feature extraction process. GoogleNet and Inception achieved accuracy scores of 64.53% and 67.09%, respectively, indicating their limitations in capturing comprehensive action details. EfficientNetB3 and Xception performed even lower, with scores of 50.70% and 52.46%, respectively, likely due to their inefficiency in handling complex video data. The 2-stream Inceptionv1 model showed moderate improvement with 66.40% but still fell short. ResNet101 demonstrated relatively high performance with 86.14% accuracy. Using a combination of CNN and Bi-LSTM architectures, ARNet effectively captures complex motion patterns and dependencies in video frames, leading to a significant boost in accuracy, achieving a score of 93.82%, the highest among the compared methods. This comprehensive analysis highlights the robustness and efficiency of ARNet in human action recognition tasks.

Table 5: ARNet performance comparison to DL approaches for the HMDB51 data sample

Model	Accuracy (%)
GoogleNet	64.53
Inception	67.09
EfficientNetB3	50.70
Xception	52.46
2stream-Inceptionv1	66.40
ResNet101	86.14
Proposed	93.82

The results for the KTH dataset against several latest DL approaches, such as 3D-ResNet, 3D-Densenet [56], VGG16 [39], and EfficientNet-B0 [57], and classification results are listed in Table 6. The values clearly show that the proposed approach outperformed the other DL frameworks. The 3D-ResNet model, with an accuracy of 91.20%, and the 3D-Densenet model, achieving 91.67%, struggle with efficiently capturing effective temporal features due to their reliance on 3D convolutional layers. VGG16, although achieving a high accuracy of 98%, primarily focuses on spatial feature extraction, missing out on nuanced temporal dynamics. EfficientNet-B0, with an accuracy of 97.30%, provides a balanced approach but is not specifically optimized for action recognition tasks. In contrast, the proposed ARNet, with an accuracy of 99%, uses a robust architecture combining CNN for spatial feature extraction and Bi-LSTM for capturing temporal dependencies. This dual approach allows ARNet to more effectively process complex motion patterns and subtle temporal variations, leading to superior performance in action recognition.

Table 6: ARNet performance comparison to DL approaches for the KTH data sample

Model	Accuracy (%)
3D-ResNet	91.20
3D-Densene	91.67

(Continued)

Table 6 (continued)

Model	Accuracy (%)
VGG16	98.00
EfficientNet-B0	97.30
Proposed	99.00

In the third experiment, the results for the UCF Sports dataset against numerous DL approaches were discussed. Hence, models such as GoogleNet and VGG16, as given in [58], VGG19 [59], and Inception-ResNetv2 [60], and the results are detailed in Table 7. The values show that the ARNet approach attains the highest recognition results. GoogleNet and VGG16, with accuracies of 74.45% and 74.64%, respectively, are limited by their less advanced temporal processing capabilities. VGG19 (97.13%) and Inception-ResNetv2 (92.90%) perform better but still struggle with the complexity of dynamic sports actions. ARNet, on the other hand, excels by integrating CNNs for detailed spatial feature extraction with Bi-LSTM networks for effective temporal analysis, leading to a superior accuracy of 99.16%. Based on all the comparative analyses performed on 3 datasets, the proposed approach is more proficient for HAR from visual samples.

Table 7: ARNet performance comparison to DL approaches for the UCF Sports data sample

Model	Accuracy (%)
GoogleNet	74.45
VGG16	74.64
VGG19	97.13
Inception-ResNetv2	92.90
Proposed	99.16

4.6 Comparison of the ARNet Approach with New Works

This section discusses the results of the proposed approach against the latest works over all three datasets named HMDB51, UCF Sports, and KTH, and the obtained evaluation is presented in Table 8.

Table 8: ARNet comparison to new works

Reference	Year	Accuracy (%)
HMDB51 Dataset		
[61]	2024	92.70
[62]	2024	77.29
[8]	2024	93.70
[63]	2023	73.12
[64]	2023	72.60
Proposed	2024	93.82
UCF Sports Dataset		
[65]	2024	93.30
[8]	2024	98.00

(Continued)

Table 8 (continued)

Reference	Year	Accuracy (%)
[66]	2023	97.84
[67]	2023	90.00
[58]	2023	99.00
Proposed	2024	99.16
	KTH Dataset	
[31]	2024	98.00
[68]	2024	90.00
[8]	2024	97.00
[67]	2023	94.00
[58]	2023	98.70
Proposed	2024	99.00

First, the results attained for the HMDB51 dataset were compared to several works [8,61–63], and the obtained analysis in terms of accuracy score is provided in the first half of Table 8. The values clearly show that our method performed better than the nominated approaches. Uddin et al. [61] proposed a DL approach using CNN, ConvLSTM, and long-term recurrent convolutional network (LRCN) architectures to extract spatial and temporal features from video data and reported the best results attained by the CNN model with an accuracy value of 92.70% on the HMDB51 dataset. The approach in [62] suggested a DL model using CNNs with channel attention mechanisms (CAMs) and autoencoders (AEs) for low-size and low-resolution videos. The work enhanced the computation of keypoints and used random frame sampling to improve accuracy. The work attained an accurate value of 77.29% on the discussed dataset. Khan et al. [8] employed VGG19 to calculate dense keypoints and extract multiview keypoints from horizontal and vertical gradients. Combined features were filtered using relative entropy, mutual information, and strong correlation coefficient (SCC). The optimal features were then utilized to train the NB predictor for label estimation and attained an accurate score of 93.70%. Further, the approach in [63] combined CNN and RNN methods for HAR in videos. It preprocessed video frames and used a fusion of CNNs for feature extraction. Extracted features were fed into a deep gated recurrent unit (GRU) network to capture temporal dependencies, followed by classification using a SoftMax layer, and reported an accuracy value of 73.12% over the HMDB51 dataset.

Chen et al. [64] developed an action recognition method using improved residual CNNs with spatial attention modules. A two-level attention mechanism was also introduced to highlight important frames and spatial regions, enhancing feature extraction across temporal and spatial dimensions, and reported an accuracy of 72.60%. In comparison, the proposed ARNet attained the highest accuracy value. ARNet distinguishes itself from comparative approaches by addressing key limitations in action recognition methodologies. Previous methods, such as those by Uddin et al. [61] and Dominic et al. [42], often struggle with integrating spatial and temporal features effectively. ARNet overcomes this challenge using advanced CNN architecture such as refined Inception-ResNet-V2 alongside optimized Bi-LSTM modules, ensuring comprehensive feature extraction and integration across video frames. Unlike approaches relying on complex architectures and attention mechanisms discussed in [62], which incurred high computational costs and inefficiencies, ARNet prioritizes computational efficiency without compromising accuracy. Further, the approaches utilizing either spatial or both spatiotemporal features, as given in [8,63], enhanced sequential modeling, still face model overfitting issues. Comparatively, the proposed ARNet overcomes this aspect by employing its refined Bi-LSTM setup, which is skillful in extracting complicated temporal information vital

for precise HAR. The ARNet approach acquired a classification accuracy of 93.82% over the HMDB51 sample, representing its effectiveness for HAR under complex scenarios.

Next, the comparative analysis of the ARNet is accomplished with the works [8,58,65–67] for the UCF Sports dataset by comparing the works in terms of accuracy score, and the comparison is given in the second half of Table 8. The discussed accuracy scores over the UCF Sports dataset prove the effectiveness of the proposed approach. Hassan et al. [65] employed a spatiotemporal approach for HAR from the video sequences. For this, the work employed MobileNetV2 CNNs for feature extraction and a BiLSTM network for capturing dependencies and processing data. The work reported a classification accuracy of 93.30% over the UCF Sport data sample. Khan et al. [8] discussed another approach that utilized a DL model for HAR and reported an accuracy score of 98%. Xiao et al. [66] integrated hierarchical feature maps with a multi-scale deformable attention module, effectively capturing spatial deformations and temporal fluctuations in video frames to detect diverse sports behaviors, and reported an accuracy of 97.84%. Palaniapan et al. [67] integrated an Encoder-Decoder Network for sample information computation, an Improved Scale-Invariant Feature Transform (iSIFT) to reduce redundancy, Quadratic Discriminant Analysis for feature optimization, and a Weighted Fusion strategy to merge essential keypoints information. The approach achieved an accuracy of 90%. Sowmyayani et al. [58] proposed an approach, spatio-temporal human action recognition network (STHARNet), an architecture for HAR integrated into a content-based video retrieval (CBVR) system. It worked by nominating keyframes through the Adaptive Genetic Optimization Procedure (AGOP) approach based on scene cuts, extracting spatial keypoints from these frames, generating Motion Energy Images (MEI) for temporal features from each GOP, and fusing these for HAR in STHARNet. The work attained an accurate score of 99% on the employed data sample. In comparison to all approaches [8,58,65–67], the proposed ARNet attained the highest accuracy results. The selected approaches have limitations that the ARNet method overcomes. Hassan et al. [65] and Xiao et al. [66] both rely heavily on CNNs and attention modules but miss finer temporal dependencies in actions. Khan et al. [8] used VGG19 and multiview keypoints but faced potential redundancy and inefficiency issues. Palaniapan et al. [67] incorporated multiple modules for feature extraction and optimization but suffered from model overfitting problems. Sowmyayani et al. [58] focused on keyframe extraction and MEI for temporal features did not fully capture all complicated patterns in action sequences. The proposed ARNet method integrates a refined Inception-ResNetv2 for robust spatial feature extraction with a Bi-LSTM for capturing temporal dependencies, optimizing performance, and reducing redundancy, model overfitting issues, thus achieving superior accuracy by effectively balancing spatial and temporal analysis.

Lastly, the study discussed the results attained for the KTH dataset with several works [31], and the obtained analysis in terms of accuracy score is provided in the last half of Table 8. Parro et al. [31] developed a DL approach for HAR that included a unit for people recognition and tracking, followed by a HAR algorithm against each located person. This algorithm preprocessed input data using a DL architecture based on LSTMs for sample information capture and classification of features with a DNN. The approach has reported an accuracy rate of 98% over the KTH dataset. Further, Memon et al. [68] trained a VGG19-based CNN-RNN DL network using transfer learning for HAR in the visual samples and reported a classification score of 90%. Khan et al. [8] used VGG19 to calculate dense keypoints, extract multiview keypoints, and filter combined features using relative entropy, mutual information, and SCC before training a Naïve Bayes predictor for label estimation and attained an accuracy of 97%. The work in [67] utilized an Encoder-Decoder framework for sample information computation, iSIFT to reduce redundancy, Quadratic Discriminant Analysis for feature optimization, and a Weighted Fusion strategy to merge keypoint information, and attained an accuracy of 94%. Sowmyayani et al. [58] proposed STHARNet, a HAR architecture using keyframes selected via AGOP, spatial keypoints, and Motion Energy Images for temporal features, integrated into a CBVR system, and

reported an accuracy score of 98.70%. In comparison, the ARNet achieved the highest classification score of 99%. The compared approaches exhibit several limitations: Parro et al. [31] employed a unit for people recognition and tracking, followed by an LSTM-based architecture for feature extraction, which is training-intensive and susceptible to overfitting, limiting its scalability. The VGG19-based CNN-RNN model in [68] achieved only 90% accuracy, indicating potential inefficiencies in feature extraction and sequence learning. Khan et al. [8] used dense keypoints and multiview gradients but faced challenges in optimizing feature selection, achieving slightly lower accuracy. The Encoder-Decoder and iSIFT method in [67] effectively reduced redundancy but still fell short in overall performance. Sowmyayani et al. [58] introduced a complex keyframe selection and MEI generation process that, while accurate, adds to computational overhead. ARNet overcomes these limitations by integrating an optimized CNN and Bi-LSTM architecture that efficiently captures both spatial and temporal features, enhancing accuracy and robustness while maintaining computational efficiency, resulting in the highest accuracy of 99%.

4.7 Cross-Dataset Evaluation

This section executes the cross-corpus evaluation, where the model is trained and tested on different datasets. Such an evaluation assists the reader in understanding the generalization power of an approach. Hence, this study accomplishes various experiments by considering the common classes of all three employed datasets. There exists a total of 3 common classes between the HMDB51 and KTH datasets, which are Walking, Running, and Jogging, while there are six common classes between the HMDB51 and UCF Sports datasets, which are Diving, Golf Swing, Kicking, Riding Horse, Swinging, and Skateboarding. Whereas the KTH and UCF Sports datasets contain no common classes, therefore, this study accomplishes four types of cross-dataset evaluations, which are as follows: (i) The model is trained on the mentioned three common classes from the HMDB51 and tested on the KTH sample, (ii) the approach takes the samples from the KTH dataset for model training while the test visuals are taken from the HMDB51, (iii) the model is trained on the mentioned six common classes from the HMDB51, and tested on the UCF Sports sample, and (iv) the model is trained on the UCF Sports samples and evaluated on the HMDB51 dataset. The results of all four experiments in terms of the accuracy metric are provided in Table 9, and they indicate that for the first evaluation, the approach attains an accuracy value of 63.09%, which is 62.28% for the second evaluation. In addition, the ARNet achieves accuracy scores of 68.26% and 65.11% for the third and fourth evaluations. The results indicate that the approach undergoes performance degradation in terms of cross-dataset evaluation; however, it still demonstrates robust and consistent performance across different evaluations. This resilience highlights the adaptability and effectiveness of the proposed model in handling diverse data variations, which is crucial for real-world applications where data can vary significantly. ARNet's ability to maintain a high level of accuracy and true positive rate across datasets highlights its potential as a reliable tool for action recognition, even in challenging and heterogeneous scenarios.

Table 9: Cross-corpus evaluation of the ARNet

Training corpus	Testing corpus	Accuracy (%)
HMDB51	UCF Sport	68.26
UCF Sport	HMDB51	65.11
HMDB51	KTH	63.09
KTH	HMDB51	62.28

5 Conclusion

This study introduces an approach, ARNet, which combines a refined Inception-ResNetv2-based CNN with a Bi-LSTM network, using novel PReLU activations to effectively integrate spatial and temporal information for action recognition in videos. Experimental validation on benchmark datasets HMDB51, KTH, and UCF Sports indicates ARNet's superior performance, achieving an accuracy of 93.82%, 99%, and 99.16%, respectively. These results highlight ARNet's robustness and potential across various domains, including security surveillance, healthcare monitoring, and human-computer interaction. In addition, this research conducts the cross-corpus evaluation to prove the generalization power of the indicated approach. The practical implications of these findings are significant: ARNet can be applied in real-world scenarios such as intelligent video surveillance, healthcare monitoring for the elderly or mobility-impaired individuals, and human-computer interaction systems where accurate activity recognition is essential. Despite its strengths, ARNet has some limitations, including the computational complexity associated with deep models and a potential decrease in performance with extremely low-quality or occluded video input. Future work will focus on addressing these challenges by exploring lightweight variants of ARNet, domain adaptation through transfer learning, and the inclusion of multimodal inputs such as audio or depth data to further enhance recognition accuracy in more complex environments.

Acknowledgement: The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding this work through (grant number IMSIU-DDRSP2504).

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2504).

Author Contributions: The authors confirm their contribution to the paper as follows: Conceptualization, Hussain Dawood and Marriam Nawaz; methodology, Marriam Nawaz and Tahira Nazir; software, Ali Javed and Abdul Khader Jilani Saudagar; validation, Ali Javed and Hatoon S. AlSagri; formal analysis, Hussain Dawood, Marriam Nawaz, and Tahira Nazir; investigation, Tahira Nazir, Ali Javed, and Abdul Khader Jilani Saudagar; resources, Abdul Khader Jilani Saudagar and Hatoon S. AlSagri; data curation, Hussain Dawood; writing—original draft preparation, Marriam Nawaz and Tahira Nazir; writing—review and editing, Ali Javed; visualization, Abdul Khader Jilani Saudagar; supervision, Hussain Dawood; project administration, Ali Javed; funding acquisition, Abdul Khader Jilani Saudagar and Hatoon S. AlSagri. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data is openly available in a public repository.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Sun Z, Ke Q, Rahmani H, Bennamoun M, Wang G, Liu J. Human action recognition from various data modalities: a review. IEEE Trans Pattern Anal Mach Intell. 2023;45(3):3200–25. doi:10.1109/tpami.2022.3183112.
2. Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/CVPR.2016.213.
3. Jhuang H, Gall J, Zuffi S, Schmid C, Black MJ. Towards understanding action recognition. In: Proceedings of the 2013 IEEE International Conference on Computer Vision; 2013 Dec 1–8; Sydney, NSW, Australia. doi:10.1109/ICCV.2013.396.
4. Herath S, Harandi M, Porikli F. Going deeper into action recognition: a survey. Image Vis Comput. 2017;60(2):4–21. doi:10.1016/j.imavis.2017.01.010.

5. Kong Y, Fu Y. Human action recognition and prediction: a survey. Int J Comput Vis. 2022;130(5):1366–401. doi:10.1007/s11263-022-01594-9.
6. Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 2013 IEEE International Conference on Computer Vision; 2013 Dec 1–8; Sydney, NSW, Australia. doi:10.1109/ICCV.2013.441.
7. Adewopo VA, Elsayed N, ElSayed Z, Ozer M, Abdelgawad A, Bayoumi M. A review on action recognition for accident detection in smart city transportation systems. J Electr Syst Inf Technol. 2023;10(1):57. doi:10.1186/s43067-023-00124-y.
8. Khan MA, Javed K, Khan SA, Saba T, Habib U, Khan JA, et al. Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimed Tools Appl. 2024;83(5):14885–911. doi:10.1007/s11042-020-08806-9.
9. Morshed MG, Sultana T, Alam A, Lee YK. Human action recognition: a taxonomy-based survey, updates, and opportunities. Sensors. 2023;23(4):2182. doi:10.3390/s23042182.
10. Karácsony T, Jeni LA, De la Torre F, Cunha JPS. Deep learning methods for single camera based clinical in-bed movement action recognition. Image Vis Comput. 2024;143:104928. doi:10.1016/j.imavis.2024.104928.
11. Ren B, Liu M, Ding R, Liu H. A survey on 3D skeleton-based action recognition using learning method. Cyborg Bionic Syst. 2024;5:0100. doi:10.34133/cbsystems.0100.
12. Thakur D, Dangi S, Lalwani P. A novel hybrid deep learning approach with GWO-WOA optimization technique for human activity recognition. Biomed Signal Process Control. 2025;99(3):106870. doi:10.1016/j.bspc.2024.106870.
13. Mehmood F, Guo X, Chen E, Akbar MA, Khan AA, Ullah S. Extended multi-stream temporal-attention module for skeleton-based human action recognition (HAR). Comput Hum Behav. 2025;163(4):108482. doi:10.1016/j.chb.2024.108482.
14. Wang L, Huynh DQ, Koniusz P. A comparative review of recent kinect-based action recognition algorithms. IEEE Trans Image Process. 2020;29:15–28. doi:10.1109/tip.2019.2925285.
15. Khan SI, Dawood H, Khan MA, Issa GF, Hussain A, Alnfiai MM, et al. Transition-aware human activity recognition using an ensemble deep learning framework. Comput Hum Behav. 2025;162(4):108435. doi:10.1016/j.chb.2024.108435.
16. Wu Q, Huang Q, Li X. Human action recognition based on STDMI-HOG and STjoint feature. In: Proceedings of the 2021 IEEE International Conference on Progress in Informatics and Computing (PIC); 2021 Dec 17–19; Shanghai, China. doi:10.1109/pic53636.2021.9687036.
17. Pavan M, Jyothi K. Human action recognition in still images using SIFT key points. In: Proceedings of the International Conference on Intelligent Systems and Sustainable Computing (ICISSC 2021); 2021 Sep 24–25. Hyderabad, India. doi:10.1007/978-981-19-0011-2_29.
18. Wang J, Shao Z, Huang X, Lu T, Zhang R, Lv X. Spatial-temporal pooling for action recognition in videos. Neurocomputing. 2021;451(3):265–78. doi:10.1016/j.neucom.2021.04.071.
19. Nawaz M, Masood M, Javed A, Iqbal J, Nazir T, Mehmood A, et al. Melanoma localization and classification through faster region-based convolutional neural network and SVM. Multimed Tools Appl. 2021;80(19):28953–74. doi:10.1007/s11042-021-11120-7.
20. Hassan M, Iqbal MM, Qayyum H, Nawaz M, Ali F. Sketch-based face recognition using deep learning. Webology. 2022;19(3):3790–807.
21. Gammulle H, Ahmedt-Aristizabal D, Denman S, Tychsen-Smith L, Petersson L, Fookes C. Continuous human action recognition for human-machine interaction: a review. ACM Comput Surv. 2023;55(13s):1–38. doi:10.1145/3587931.
22. Nawaz M, Javed A, Irtaza A. Convolutional long short-term memory-based approach for deepfakes detection from videos. Multimed Tools Appl. 2024;83(6):16977–7000. doi:10.1007/s11042-023-16196-x.
23. Nawaz M, Javed A, Irtaza A. A deep learning model for FaceSwap and face-reenactment deepfakes detection. Appl Soft Comput. 2024;162(6):111854. doi:10.1016/j.asoc.2024.111854.
24. Jia X, Li W, Wang Y, Hong SC, Su X. An action unit co-occurrence constraint 3DCNN based action unit recognition approach. KSII Trans Internet Inf Syst. 2020;14(3):924–42. doi:10.3837/tiis.2020.03.001.

25. Truong TD, Bui QH, Duong CN, Seo HS, Phung SL, Li X, et al. DirecFormer: a directed attention in transformer approach to robust action recognition. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.01940.
26. Bhola G, Vishwakarma DK. A review of vision-based indoor HAR: state-of-the-art, challenges, and future prospects. Multimed Tools Appl. 2024;83(1):1965–2005. doi:10.1007/s11042-023-15443-5.
27. Moutik O, Sekkat H, Tigani S, Chehri A, Saadane R, Tchakoucht TA, et al. Convolutional neural networks or vision transformers: who will win the race for action recognitions in visual data? Sensors. 2023;23(2):734. doi:10.3390/s23020734.
28. Garcia-Gonzalez D, Rivero D, Fernandez-Blanco E, Luaces MR. New machine learning approaches for real-life human activity recognition using smartphone sensor-based data. Knowl Based Syst. 2023;262(1):110260. doi:10.1016/j.knosys.2023.110260.
29. Kapoor S, Sharma A, Verma A, Dhull V, Goyal C. A comparative study on deep learning and machine learning models for human action recognition in aerial videos. Int Arab J Inf Technol. 2023;20(4):567–74. doi:10.34028/iajit/20/4/2.
30. Azmat U, Alotaibi SS, Abdelhaq M, Alsufyani N, Shorfuzzaman M, Jalal A, et al. Aerial insights: deep learning-based human action recognition in drone imagery. IEEE Access. 2023;11:83946–61. doi:10.1109/access.2023.3302353.
31. Cob-Parro AC, Losada-Gutiérrez C, Marrón-Romera M, Gardel-Vicente A, Bravo-Muñoz I. A new framework for deep learning video based Human Action Recognition on the edge. Expert Syst Appl. 2024;238(13):122220. doi:10.1016/j.eswa.2023.122220.
32. Zhang R, Li S, Xue J, Lin F, Zhang Q, Ma X, et al. Hierarchical action recognition: a contrastive video-language approach with hierarchical interactions. arXiv:2405.17729v1. 2024.
33. Mekruksavanich S, Jitpattanakul A. Device position-independent human activity recognition with wearable sensors using deep neural networks. Appl Sci. 2024;14(5):2107. doi:10.3390/app14052107.
34. Khan D, Alonazi M, Abdelhaq M, Al Mudawi N, Algarni A, Jalal A, et al. Robust human locomotion and localization activity recognition over multisensory. Front Physiol. 2024;15:1344887. doi:10.3389/fphys.2024.1344887.
35. Kaya Y, Topuz EK. Human activity recognition from multiple sensors data using deep CNNs. Multimed Tools Appl. 2024;83(4):10815–38. doi:10.1007/s11042-023-15830-y.
36. Kolkar R, Geetha V. Human activity recognition using deep learning techniques with spider monkey optimization. Multimed Tools Appl. 2023;82(30):47253–70. doi:10.1007/s11042-023-15007-7.
37. Brishtel I, Krauss S, Chamseddine M, Rambach JR, Stricker D. Driving activity recognition using UWB radar and deep neural networks. Sensors. 2023;23(2):818. doi:10.3390/s23020818.
38. Surek GAS, Seman LO, Stefenon SF, Mariani VC, Coelho LDS. Video-based human activity recognition using deep learning approaches. Sensors. 2023;23(14):6384. doi:10.3390/s23146384.
39. Alhakbani N, Alghamdi M, Al-Nafjan A. Design and development of an imitation detection system for human action recognition using deep learning. Sensors. 2023;23(24):9889. doi:10.3390/s23249889.
40. Dwivedi N, Singh DK, Kushwaha DS. A novel approach for suspicious activity detection with deep learning. Multimed Tools Appl. 2023;82(21):32397–420. doi:10.1007/s11042-023-14445-7.
41. Gowada R, Pawar D, Barman B. Unethical human action recognition using deep learning based hybrid model for video forensics. Multimed Tools Appl. 2023;82(19):28713–38. doi:10.1007/s11042-023-14508-9.
42. Dominic N, Cenggoro TW, Budiarto A, Pardamean B. Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification. Commun Math Biol Neurosci. 2021;2021:39. doi:10.28919/cmbn/5565.
43. Wang Z, Yin Y, Yin R. Multi-tasking atrous convolutional neural network for machinery fault identification. Int J Adv Manuf Technol. 2023;124(11):4183–91. doi:10.1007/s00170-022-09367-x.
44. Suebsombut P, Sekhari A, Sureephong P, Belhi A, Bouras A. Field data forecasting using LSTM and Bi-LSTM approaches. Appl Sci. 2021;11(24):11820. doi:10.3390/app112411820.
45. Su C, Wang W. Concrete cracks detection using Convolutional Neural Network based on transfer learning. Math Probl Eng. 2020;2020(13):7240129. doi:10.1155/2020/7240129.

46. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: a large video database for human motion recognition. In: Proceedings of the 2011 International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain. doi:10.1109/ICCV.2011.6126543.
47. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB51 [Internet]. [cited 2025 Jun 15]. Available from: https://paperswithcode.com/dataset/hmdb51.
48. Rodriguez MD, Ahmed J, Shah M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition; 2008 Jun 23–28; Anchorage, AK, USA. doi:10.1109/CVPR.2008.4587727.
49. UCF Sports Action Data Set [Internet]. [cited 2025 Jun 15]. Available from: https://www.crcv.ucf.edu/data/UCF_Sports_Action.php.
50. Liu J, Luo J, Shah M. Recognizing realistic actions from videos in the wild. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. doi:10.1109/CVPR.2009.5206744.
51. KTH dataset [Internet]. [cited 2025 Jun 15]. Available from: https://www.csc.kth.se/cvap/actions/.
52. Jahandad, Sam SM, Kamardin K, Amir Sjarif NN, Mohamed N. Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3. Procedia Comput Sci. 2019;161(8):475–83. doi:10.1016/j.procs.2019.11.147.
53. Muneer Butt U, Aman Ullah H, Letchmunan S, Tariq I, Hafinaz Hassan F, Wei Koh T. Leveraging transfer learning for spatio-temporal human activity recognition from video sequences. Comput Mater Contin. 2023;74(3):5017–33. doi:10.32604/cmc.2023.035512.
54. Kim JH, Won CS. Action recognition in videos using pre-trained 2D convolutional neural networks. IEEE Access. 2020;8:60179–88. doi:10.1109/ACCESS.2020.2983427.
55. Silva D, Manzo-Martínez A, Gaxiola F, Gonzalez-Gurrola L, Ramírez-Alonso G. Analysis of CNN architectures for human action recognition in video. Computación y Sistemas. 2022;26(2):623–41. doi:10.13053/cys-26-2-4245.
56. Park JH, Lee EJ. Human activity recognition based on 3D residual dense network. J Korea Multimed Soc. 2020;23(12):1540–51. doi:10.9717/kmms.2020.23.12.1540.
57. Naeem Akbar M, Khan S, Umar Farooq M, Alhaisoni M, Tariq U, Usman Akram M. HybridHR-Net: action recognition in video sequences using optimal deep learning fusion assisted framework. Comput Mater Contin. 2023;76(3):3275–95. doi:10.32604/cmc.2023.039289.
58. Sowmyayani S, Rani PAJ. STHARNet: spatio-temporal human action recognition network in content based video retrieval. Multimed Tools Appl. 2023;82(24):38051–66. doi:10.1007/s11042-022-14056-8.
59. Verma KK, Mohan Singh B. Vision based human activity recognition using deep transfer learning and support vector machine. In: Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON); 2021 Nov 11–13; Dehradun, India. doi:10.1109/UPCON52273.2021.9667661.
60. Uddin MA, Lee YK. Feature fusion of deep spatial features and handcrafted spatiotemporal features for human action recognition. Sensors. 2019;19(7):1599. doi:10.3390/s19071599.
61. Uddin MA, Talukder MA, Uzzaman MS, Debnath C, Chanda M, Paul S, et al. Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN. Int J Cogn Comput Eng. 2024;5(3):259–68. doi:10.1016/j.ijcce.2024.06.004.
62. Dastbaravardeh E, Askarpour S, Saberi Anari M, Rezaee K. Channel attention-based approach with autoencoder network for human action recognition in low-resolution frames. Int J Intell Syst. 2024;2024:1052344. doi:10.1155/2024/1052344.
63. Abdelrazik MA, Zekry A, Mohamed WA. Efficient hybrid algorithm for human action recognition. J Image Graph. 2023;11(1):72–81. doi:10.18178/joig.11.1.72-81.
64. Chen B, Meng F, Tang H, Tong G. Two-level attention module based on spurious-3D residual networks for human action recognition. Sensors. 2023;23(3):1707. doi:10.3390/s23031707.
65. Hassan N, Miah ASM, Shin J. A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. Appl Sci. 2024;14(2):603. doi:10.3390/app14020603.

66. Xiao L, Cao Y, Gai Y, Khezri E, Liu J, Yang M. Recognizing sports activities from video frames using deformable convolution and adaptive multiscale features. J Cloud Comput. 2023;12(1):167. doi:10.1186/s13677-023-00552-1.
67. Palaniapan SG, Sok Choo AN. A hybrid model for human action recognition based on local semantic features. J Adv Res Comput Appl. 2024;33(1):7–21. doi:10.37934/arca.33.1.721.
68. Memon FA, Memon MH, Ali Halepoto I, Memon R, Bhangwar AR. Action recognition in videos using VGG19 pre-trained based CNN-RNN deep learning model. VFAST Trans Softw Eng. 2024;12(1):46–57. doi:10.21015/vtse.v12i1.1711.