IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 350~360
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp350-360

Human activity recognition using selective kernel network-2D convolutional neural network with ArcFace loss

Banushri Srinivasaiah, Jagadeesha Ramegowda
Department of Computer Science and Engineering, Impact College of Engineering and Applied Sciences, Bengaluru, India
Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Mar 7, 2025
Revised Nov 3, 2025
Accepted Jan 10, 2026

Keywords:

2D convolutional neural network
ArcFace loss
Bidirectional gated recurrent unit
Human activity recognition
Selective kernel network

ABSTRACT

Human activity recognition (HAR) is a widely adopted technique in applications requiring accurate identification of human actions. However, HAR approaches often face challenges in generalizing across complex datasets with multi-view variations, resulting in reduced classification accuracy. Existing classifiers face shortcomings in predicting human activities due to the presence of irrelevant video frames, leading to frequent misclassifications. This research proposes a selective kernel network-2D convolutional neural network with additive angular margin loss for deep face recognition (SKN-2D-CNN with ArcFace loss) to recognize human activity effectively. SKN dynamically adapts the receptive field for learning multi-scale spatial features, enhancing the recognition of intricate human activities with varying motion scales. In the embedding space, ArcFace loss introduces an angular margin penalty that improves inter-class separability and intra-class compactness for recognition. Together, the proposed method minimizes misclassification in human activity by improving the robustness of feature representation. Feature extraction using visual geometry group 19 (VGG19) captures spatial features like edges, textures and shapes from video frames, enhancing the model's ability to differentiate between complex human activities. The proposed method achieves high accuracy of 99.16 and 98.75% on the UCF101 and HMDB-51 datasets, respectively, when compared with existing methods such as CNN and bidirectional gated recurrent unit (BiGRU).

Corresponding Author:

Banushri Srinivasaiah
Department of Computer Science and Engineering, Impact College of Engineering and Applied Sciences
Sahakarnagar, Bengaluru-560092, India
Email: banushri914@gmail.com

1. INTRODUCTION

Human activity recognition (HAR) has emerged as an active research area due to its vital role in video understanding [1], [2]. The HAR process also detects activities based on anomalies in daily routines, which are often physically linked to a person's psychological state and personality [3]. Leveraging these action-specific characteristics enhances both the efficiency and accuracy of the activity recognition process. Furthermore, the human body's joints are hierarchically organized, providing a structured connectivity that is effectively exploited using deep learning (DL) techniques [4]–[6]. Video-based and sensor-based technologies represent the two primary approaches to activity recognition systems. Video-based systems process visual data from camera footage to recognize human activities, while sensor-based systems use external sensors to capture mobility data and monitor activity patterns [7].

DL techniques have outperformed traditional handcrafted feature-based methods, demonstrating significant success across various computer vision tasks [8]. The self-learning capability of DL networks enables them to process complex representations of visual data, making them particularly suitable for video-based HAR [9], [10]. Traditional machine learning approaches require steps such as feature extraction and selection for training, whereas modern DL models utilize kernel-based filters to process data through convolution operations, automatically extracting relevant features [11], [12]. However, external conditions such as lighting variations, camera positioning and subject distance can negatively impact recognition performance [13]. Advanced DL approaches capture spatial dependencies by learning hierarchical feature representations through convolutional processes that model both global and local spatial correlations between pixels. Additionally, neural-controlled differential equation networks are capable of capturing integral expressions of human activity, contributing to improved modeling [14], [15]. Recognizing visually similar activities remains a significant challenge due to subtle variations in human actions that represent different behaviors. Accurate differentiation is critical for reliable decision-making in applications such as surveillance, whereas misclassifying similar actions may result in overlooking abnormal or risky behavior.

Ahmad *et al*. [16] introduced a convolutional neural network (CNN) combined with a bidirectional gated recurrent unit (BiGRU) for HAR using visual data. The CNN was employed to extract deep features from frame sequences of human activity videos. The most significant features were selected to improve performance, and BiGRU was used to learn temporal motions across frames. This approach primarily aimed to enhance classification accuracy and effectively learning long-duration temporal actions. Sinha and Kumar [17] proposed a HAR framework focused on improving classification performance. The method involved segmenting images into smaller regions for feature extraction, where grey level co-occurrence matrix (GLCM) and local gradient threshold pattern (LGTP) were applied for feature extraction, and classifiers like BiGRU, CNN, and long short-term memory (LSTM) were utilized to achieve accurate classification.

Kushwaha *et al*. [18] designed a deep CNN based on multi-scale processing for HAR. A small micro-network was introduced to extract exclusive discriminative features of human objects such as pose, orientation, and object size. However, CNNs struggle to process large, data-specific human-centric features, which can lead to overfitting when the dataset lacks diversity in poses, orientations, or object sizes. Varshney and Bakariya [19] developed a deep CNN for HAR in video sequences by integrating multiple CNN streams, including spatial and temporal components. The spatial stream extracts activity representations from RGB frames, while the temporal stream captured motion-related information. However, HAR models faced limitations in handling environmental variations and occlusions, as they relied on fixed spatial and temporal features. Ahmad and Wu [20] introduced spatial deep features incorporation using a multilayer GRU for HAR. This method extracted spatial and deep features from frame sequences of human activity videos, leveraging lightweight MobileNetV2 model. The extracted features were subsequently passed through a multilayer GRU, which processed data sequences and captured temporal dependencies across video frames. The existing classifiers encountered difficulties in accurately predicting human activities due to the presence of irrelevant video frames, leading to misclassification. In order to address this challenge, this study proposes a selective kernel network-2D convolutional neural network with ArcFace loss (SKN-2D-CNN with ArcFace loss) by incorporating a dynamic kernel selection method. In contrast to traditional CNNs, the selective kernel enabled the model to adaptively adjust its receptive field based on input frames, allowing it to focus on the most informative spatial features while suppressing irrelevant background information. Additionally, the integration of ArcFace loss enhances intra-class compactness and inter-class separability, resulting in more discriminative feature representations. This combination ensures that only the most distinctive and relevant video frames contribute to activity recognition, thereby improving the overall performance and robustness of the model compared to conventional CNN approaches.

Spatial deep feature incorporation utilizing a multilayer GRU is used for HAR. This method extracts spatial and deep features from frame sequences of human activity videos by leveraging the lightweight MobileNetV2 model. The extracted features are subsequently passed through a multilayer GRU, which processes the data sequence and captures temporal dependencies across video frames. Existing classifiers face difficulties in predicting human activities due to irrelevant video frames, leading to inaccurate classification. To address this issue, a SKN-2D-CNN with ArcFace loss is proposed by incorporating a dynamic kernel selection method. In traditional CNNs, the selective kernel enables the model to adaptively adjust its receptive field depending on the input frames, focusing on the most informative spatial features while suppressing irrelevant background. Moreover, integration of ArcFace loss enhances intra-class compactness and inter-class separability, making the learned feature representations more discriminative. This combination ensures that only the most distinctive and relevant video frames contribute to activity recognition, thereby enhancing the overall performance and robustness of the model compared to standard CNNs.

The key contribution of this research study is as follows:

i) Feature extraction using visual geometry group 19 (VGG19) extracts hierarchical features from deep layers, capturing both low and high-level data and reducing video frame sizes to match the input dimensions.
ii) In traditional 2D-CNN, SKN is incorporated to dynamically adjust the receptive field, allowing the model to focus on the most relevant spatial features by avoiding irrelevant background frames. This adaptability improves the network's ability to identify intricate human activities with complex spatial patterns, leading to enhanced recognition accuracy.
iii) The ArcFace loss function enhances discriminative ability by introducing an angular margin, which enables greater inter-class separability and intra-class compactness in the feature embedding space. This results in more accurate recognition and robust performance for human activities with subtle variations between classes.

The paper is organized as follows. Section 2 details the functioning of the proposed methodology. Section 3 presents the results and discussion, while section 4 concludes the research.

2. PROPOSED METHOD

The SKN-2D-CNN with ArcFace loss efficiently captures multi-scale features, enhancing the model's ability to handle activities with diverse spatial and temporal complexities. This network dynamically adapts to the input data and effectively processes spatial information to recognize activities while producing highly discriminative feature embeddings, thereby improving classification accuracy. Initially, data is collected from the UFC101 and HMDB-51 datasets and pre-processed through normalization and removal of unwanted data. Feature extraction using VGG19 captures spatial features like edges, textures and shapes from video frames, further enhancing the model's capability to differentiate between complex human activities. Figure 1 illustrates the pipeline of the proposed methodology.

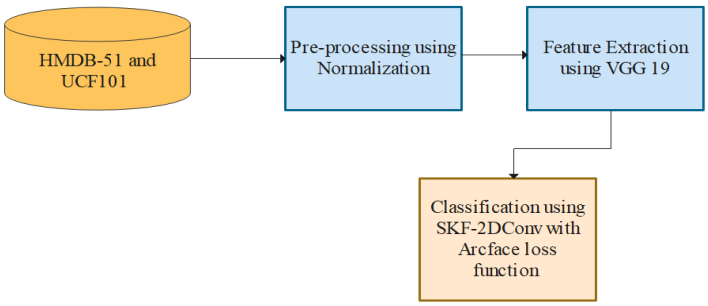


Figure 1. Pipeline of proposed method

2.1. Dataset

The video frames required for human activity classification are gathered from two publicly available datasets namely, HMDB-51 and UCF101. The UCF101 dataset includes action recognition data from realistic behavior videos with 101 activity categories, collected from YouTube. Most of the video clips in the HMDB-51 dataset are sourced from movies, with a smaller portion extracted from public platforms such as Google Video, YouTube, and the Prelinger Archive. A detailed explanation of these datasets is provided.

2.1.1. UCF-101 dataset

The UCF-101 dataset contains action recognition data from realistic behavior videos gathered from YouTube, encompassing 101 activity categories [21]. It is an extension of the UCF50 dataset, comprising 13,320 videos across these 101 categories. UCF-101 offers a wide variety of human actions with notable variations in camera motion, target appearance, target scale, and viewing angles. The 101 action categories are organized into 25 groups, each containing 4 to 7 action videos.

2.1.2. HMDB-51 dataset

Most of the video clips in the HMDB-51 dataset [22] are sourced from motion pictures, with a smaller portion obtained from public platforms like Google Video, YouTube, and the Prelinger Archive. The dataset contains a total of 6,849 clips. The activities are categorized into five categories: general body

movements, body movements involving object interaction, object-influenced facial behavior, general facial behavior, and body movements involving human interaction. Table 1 provides a description of the UCF-101 and HMDB-51 datasets used for video-based HAR, detailing the number of activities, video clips, data sources, and training-testing splits.

Table 1. Detailed description of datasets used for video based HAR

Dataset	HMDB-51	UCF-101
Number of activities	51	101
Video clips	6766	13,320
Training set	70 video clips	14,900
Sources	YouTube	YouTube, Movie clips, and Google video
Test set	30 video clips	6,360

2.2. Pre-processing

This section describes the data pre-processing phase, where a sliding window approach is considered and normalization is applied to equalize the feature vector. To avoid large-scale features from dominating the dataset, normalization is used to put all features into a comparable range. Min-max normalization significantly enhances the accuracy of a machine learning model by scaling dataset values to the [0, 1] range. In (1) represents the transformation function.

$$X_{new}^{*} = \left(\frac{X_{Old} - X_{min}}{X_{max} - X_{min}}\right) \quad (1)$$

Where $X_{Old}, X_{max}, X_{min}$ denote the original, maximum and minimum values of the given features, respectively. Where X_{new}^{*} denotes the normalized value of X_{Old}, scaled between the range [0,1]. The minimum and maximum values are extracted from the training set and used to normalize both the training and testing datasets.

i) In HAR using sensors, the sliding window technique generates time-series data, where dependencies exist between previous and recent values. An effective feature generation mechanism is essential in HAR to precisely capture temporal dependencies.
ii) The sliding window method groups sensor readings, with each window containing multiple features from the same time step. As the window moves forward by s steps, the next sample includes readings from s+1 to s+w+1, where s refers to step size and w denotes the window size.
iii) To ensure effective data handling, the test set is pre-processed before being fed for feature extraction.

2.3. Feature extraction

The VGG19 is a type of CNN consisting of 19 layers, including 16 convolution (conv) layers and 3 fully connected (FC) layers, designed to classify HAR into 1000 categories. It is trained on image data, utilizing multiple 3×3 filters in each conv layer. The 16-conv layer performs feature extraction, while the final 3 layers handle classification. Features are extracted in 5 groups, each followed by a max-pooling layer. The model receives an image of the size 224×224, with outcomes corresponding to the action recognition label [23], [24]. The VGG19 features a structured design comprising multiple connected conv layers and fully connected layers. It involves an alternative arrangement of conv layers and non-linear activation layers, outperforming single-layer structures by effectively extracting image features. Max-pooling is used for down sampling, while a modified rectified linear unit (ReLU) activation function selects the largest values within a region as pooled values. The downsampling layer enhances feature extraction by retaining key details while reducing the number of parameters. The initial 16-conv layers of VGG19 extract features, while the subsequent two layers classify them. Each feature extraction group is separated by a max-pooling layer. After feature extraction, data is passed to classification layers for HAR identification.

2.4. Classification

This section introduces the SKN-2D-CNN with ArcFace loss function, which efficiently captures multi-scale features, allowing the model to handle activities with varying spatial and temporal complexities. In traditional 2D-CNN, SKN is incorporated to enable the network to dynamically adjust its receptive field based on input features through a split, fuse, and select process. In the split stage, input feature maps are convolved using multiple kernels of different sizes, such as 3×3 and 5×5, to capture diverse spatial patterns. During the fuse stage, these outputs are integrated using global average pooling and fully connected layers to produce compact, channel-wise attention descriptors. In the select stage, a soft attention mechanism is applied to select the most appropriate features across the different kernel scales, helping capture fine-grained information that distinguishes intricate human activities. The ArcFace loss function is integrated into the classification phase to enhance the discriminative ability of the model [25]. Instead of using Euclidean

distance, ArcFace loss introduces an angular margin penalty between feature embeddings of different classes. This margin improves inter-class separability and intra-class compactness, resulting in more consistent and robust feature learning. Together, the proposed method ensures high recognition performance by focusing on the most relevant spatial features and generating highly separable representations in the embedding space. Initially, the setup uses 2 kernels of various sizes but is easily extendable to multiple branches. The automatically initialized split feature map $X \in \mathbb{R}^{H'} \times W' \times C'$ conducts 2 transformations $\tilde{F}: X \to \tilde{U} \in \mathbb{R}^{H \times W \times C}$ using kernels of size 3×3 and 5×5. Both $\tilde{F}$ and $\hat{F}$ have efficiently grouped convolution and batch normalization in sequence. The fuse operation allows neurons for adjusting RF sizes based on the context of the input stimuli. The core idea is to utilize gating mechanisms to control the flow of data from several branches, each carrying information at different scales, into the neurons of subsequent layers. The initial fusion of outputs from the branches is performed element-wise, as defined in (2). Here, global information is embedded by applying global average pooling for generating channel-wise statistics, represented as $s \in \mathbb{R}^C$. The $c - th$ element of s is evaluated by reducing U across the spatial dimension $H \times W$, as expressed by (3).

$$U = \tilde{U} + \hat{U} \tag{2}$$

$$s_c = F_{gp}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c\,(i,j) \tag{3}$$

Where U determines feature maps, $\tilde{U} + \hat{U}$ refers to feature maps from two branches of different kernel sizes, c represents a channel, H and W denote the height and weight of spatial dimensions, and $U_c(i,j)$ denotes the activation value at spatial location (i,j) in c channel of the feature map. Additionally, a compact feature $z \epsilon\, \mathbb{R}^{d \times 1}$ is constructed for enabling adaptive and precise selections using FC layer with dimensionality reduction, improving efficiency. The ReLU function is represented by β and $W \epsilon \mathbb{R}^{d \times 1}$. To evaluate the impact of d on model effectiveness, a decreased ratio r is introduced to regulate its values as expressed in (4) to (5). Where L indicates the minimum distance d typically set to 32. A soft attention mechanism is utilized through channels for the selection of several spatial data scales using feature z. The SoftMax is used for channel-wise process, as represented in (6) and (7).

$$z = \mathcal{F}_{f_c}(s) = \delta\big(\beta(W_s)\big) \tag{4}$$

$$d = max\left(\frac{C}{r}, L\right) \tag{5}$$

$$a_c = \frac{e^{A_c z}}{e^{A_c z} + e^{B_c z}}, b_c = \frac{e^{B_c z}}{e^{A_c z} + e^{B_c z}} \tag{6}$$

$$V_c = a_c . \tilde{U}_c + b_c . \hat{U}_c, a_c + b_c = 1 \tag{7}$$

Where e^A, e^B denotes the attention weight for channel c in branches A and B, and e represents Euler's number. Each block has 2 conv layers, where the initial layer employs 1×1 kernels to enhance feature representation, while the second layer uses user kernels of size 3 to aggregate consecutive information and extract increasingly complicated relations. Each conv layer is followed using batch normalization and a ReLU activation. After each 2 conv blocks, a FC layer with dropout is applied to generate sequential feature vector $v_{(u,t)} \in \mathbb{R}^d$. The concatenated sequential vector $v_{(u,t)}$ with user embedding e_u then connects the outcomes layer with $|I|$ node to the loss function.

The 2D-CNN with ArcFace function ensures that every class eliminates irrelevant features, enhancing discriminative ability of learned embedding feature during classification. The 2D-CNN model evaluates similarity between the input feature embeddings and all classes is evaluated utilizing cosine similarity score for prediction. ArcFace loss function introduces an angular margin to the cosine similarity score of the ground truth class, which increases angular separation among different class embeddings. This margin is incorporated into the log-SoftMax formulation to normalize similarity scores, producing a valid probability distribution across all classes. By focusing on angular relationships rather than the absolute magnitudes of embeddings, ArcFace loss simplifies the optimization process by leveraging cosine similarity scores, which are inherently bounded between -1 and 1, making training more effective and stable. During training, the model learns to separate embeddings of several classes in angular space, resulting in predictions that are predominantly distributed between 0 and 1. This approach improves the reliability of the fuzzy logic system as expressed in (8). Where T represents the temperature, which regulates the probability distribution

at output, y_i denotes input for i^{th} class, and N determines the total number of classes. The higher values of T make the output layer distribution closer to a uniform distribution.

$$Softmax(x)_i = \frac{e^{y_i/T}}{\sum_{i=1}^{N} e^{y_i/T}} \quad (8)$$

3. EXPERIMENTAL ANALYSIS

This section presents an evaluation of the proposed 2D-CNN with ArcFace loss function simulated using Python version 3.11, Windows 10 operating system (OS), 6 GB GPU, 16 GB RAM, i7 processor, and 1 GB memory. The proposed methodology is assessed based on the performance measures of F1-score, accuracy, recall, and precision, as shown in (9) to (12). Where TN denotes true negative, TP denotes true positive, FN denotes false negative, and FP denotes false positive.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FP} \quad (11)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (12)$$

3.1. Performance analysis

Table 2 presents an evaluation of different feature extraction methods across both datasets. The performance analysis of VGG19 is proven more efficient in capturing complex patterns and extracting hierarchical feature base on the dataset. The feature extraction performance of the proposed model is compared with existing methods like EfficientNet, MobileNet, ResNet50, and VGG16. The suggested model achieves a better accuracy of 99.16 and 98.75% on the UFC101 and HMDB-51 datasets, respectively.

Table 2. Performance evaluation of different feature extraction methods on both dataset

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNet	UFC101	92.14	91.67	90.82	91.24
	HMDB-51	89.24	88.76	87.58	88.17
MobileNet	UFC101	90.67	90.21	89.15	89.68
	HMDB-51	86.53	85.89	85.23	85.56
ResNet50	UFC101	94.25	93.87	92.96	93.41
	HMDB-51	90.56	89.98	89.12	89.54
VGG16	UFC101	93.74	93.12	92.25	92.68
	HMDB-51	88.94	88.32	87.76	88.04
Proposed VGG19	UFC101	99.16	99.10	99.11	99.10
	HMDB-51	98.75	99.41	96.03	97.54

The performance analysis of the classification using SKN-2D-CNN is compared with existing methods such as multi layer perceptron (MLP), deep neural network (DNN), ResNet, and VGG. Classification using SKN-2D-CNN achieves higher accuracy of 99.16 and 98.75% on the UCF101 and HMDB-51 datasets, respectively, as shown in Table 3. The ArcFace loss outperforms traditional loss functions by ensuring that embeddings are both well-separated across classes and compact within each class. Table 4 presents the results of the ArcFace loss function based on each dataset.

Table 3. Performance analysis of different classification methods

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	UFC101	84.37	82.56	81.78	82.16
	HMDB-51	80.24	79.45	78.36	78.9
DNN	UFC101	91.52	90.78	89.34	90.05
	HMDB-51	87.96	87.42	86.34	86.87
ResNet	UFC101	93.42	92.86	91.78	92.32
	HMDB-51	89.74	89.15	88.42	88.78
VGG	UFC101	92.81	92.12	91.06	91.58
	HMDB-51	88.45	87.89	87.23	87.55
SKN-2D-CNN	UFC101	99.16	99.10	99.11	99.10
	HMDB-51	98.75	99.41	96.03	97.54

Table 4. Performance evaluation of different loss function

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Huber loss	UFC101	91.82	91.23	90.54	90.88
	HMDB-51	87.34	86.85	85.73	86.28
Pinball loss	UFC101	89.56	89.01	87.85	88.42
	HMDB-51	85.92	85.37	84.23	84.79
Categorical cross entropy	UFC101	93.54	92.98	91.85	92.41
	HMDB-51	88.67	88.09	87.25	87.66
Focal loss	UFC101	94.83	94.37	93.24	93.8
	HMDB-51	90.24	89.65	89.02	89.33
Arcface loss	UFC101	99.16	99.10	99.11	99.10
	HMDB-51	98.75	99.41	96.03	97.54

Figure 2 represents the accuracy vs epoch graph, demonstrating model's accuracy over training epochs for training and validation on both datasets: Figure 2(a) for UCF101 and Figure 2(b) for HMDB-51. The training and validation accuracy increase steadily, thereby eliminating the risk of underfitting. Figure 3 shows the loss vs. epoch graphs for proposed method, with Figure 3(a) representing UCF101 and Figure 3(b) representing HMDB-51, illustrating how the model's loss changes over the training epochs and how closely the model's predictions match the true labels. The ROC curve graphically illustrates a classifier's performance by plotting the true positive rate (TPR) against the false positive rate (FPR) across different threshold settings.

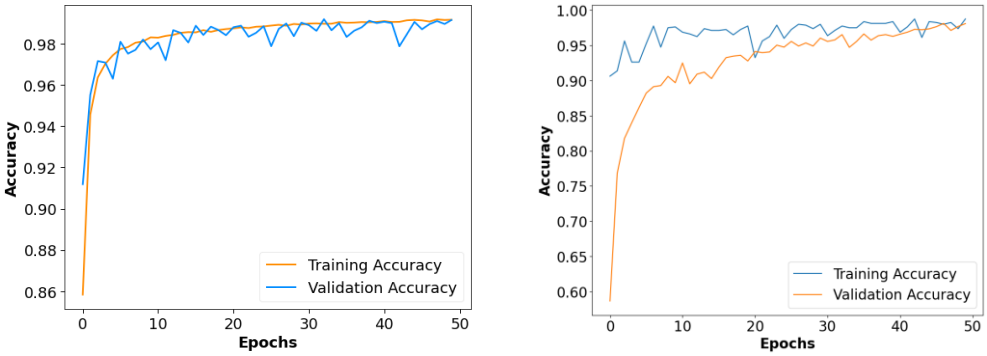


Figure 2. Performance analysis of accuracy vs epochs for proposed method of (a) UCF101 (b) HMDB-51

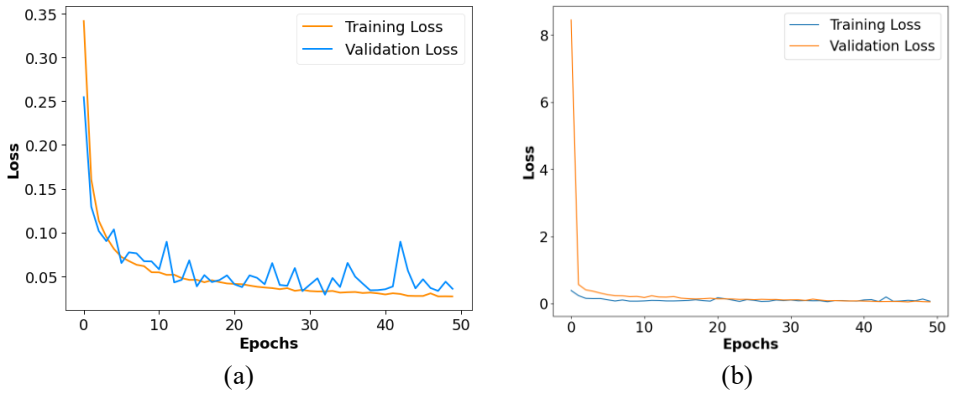
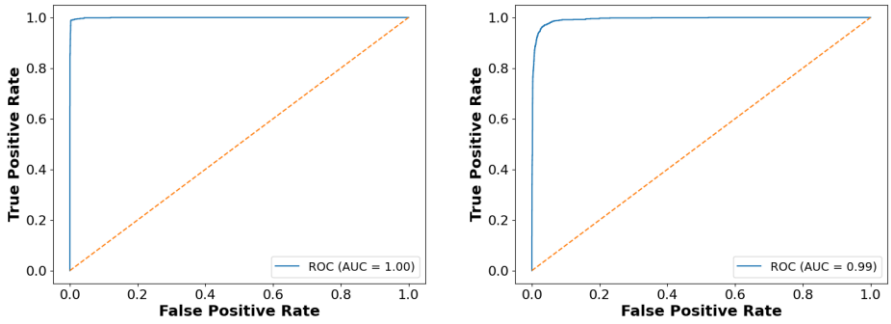


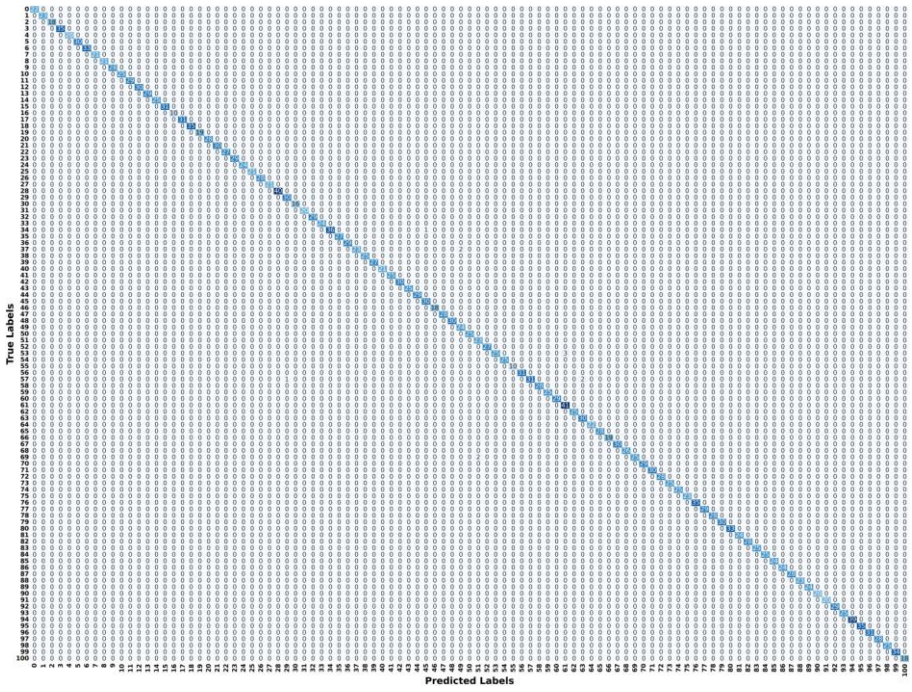
Figure 3. Performance analysis of loss vs epoch for proposed method of (a) UCF101 (b) HMDB-51

Figure 4 presents the performance analysis based on area under the curve (AUC) measures for UCF101 as shown in Figure 4(a) and HMDB-51 as shown in Figure 4(b). Higher AUC values, closer to 1, indicate better class discrimination. The performance analysis using the confusion matrix is shown in Figure 5, with Figure 5(a) for UCF101 and Figure 5(b) for HMDB-51, respectively. In each matrix, rows represent the actual classes, while columns represent the predicted classes, helping to identify misclassifications across different activity categories.

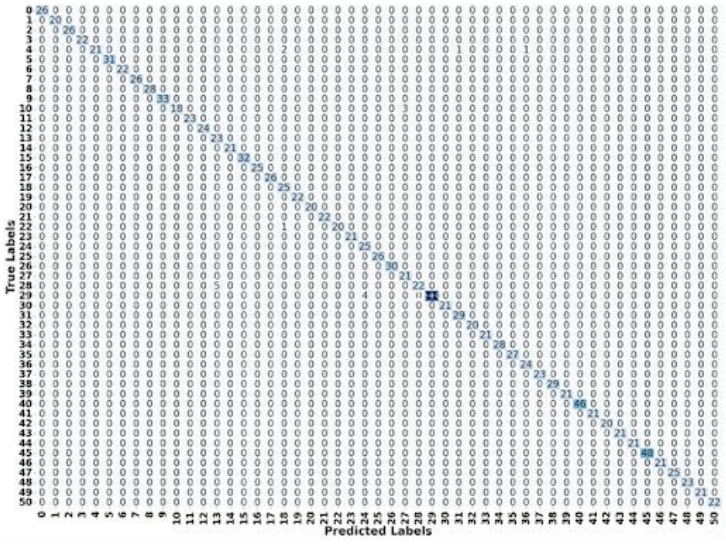


(a) (b)

Figure 4. Performance analysis of TPR vs FPR for proposed method of (a) UCF101 (b) HMDB-51



(a)



(b)

Figure 5. Performance analysis of confusion matrix for proposed method of (a) UCF101 (b) HMDB-51

3.2. Comparative analysis

Table 5 demonstrates the comparative analysis of the proposed method with existing methods. The proposed SKN-2D-CNN with Arcface loss method is compared with existing methods: CNN-BiGRU [16], BiGRU, CNN and LSTM [17], CNN [18], and MobileNetV2 [20]. The proposed SKN-2D-CNN with Arcface loss demonstrates a superior accuracy of 99.16 and 98.75% on UFC101 and HMDB-51 datasets. The SKN allows the model to adapt to diverse spatial patterns, improving recognition for complex activities and arcface loss, ensuring that feature embedding reduces misclassification for activities with high inter-class similarity.

Table 5. Comparative analysis of proposed method with existing methods

Methods	Dataset	Accuracy (%)
CNN-BiGRU [16]	UFC101	91.79
	HMDB-51	71.89
BiGRU, CNN and LSTM [17]	UFC101	98.80
	HMDB-51	NA
CNN [18]	UFC101	98.01
	HMDB-51	97.45
MobileNetV2 [20]	UFC101	92.93
	HMDB-51	80.61
Proposed SKN-2D-CNN with ArcFace loss method	UFC101	99.16
	HMDB-51	98.75

3.3. Discussion

The merits of the proposed SKN-2D-CNN with ArcFace loss and the limitations of existing techniques like HAR, which mainly focus on improving the accuracy and learning long-sequence temporal actions are discussed in this section. The HAR process typically works on GPUs but struggles to predict activities on internet of things (IoT) devices. GLCM and LGTP are used for feature extraction, while classifiers like CNN, BiGRU, and LSTM offer accurate classification. The SoftMax classifier is utilized for activity classification. The reliance on specific human-centric features leads to overfitting if the dataset lacks diversity in poses, orientations, or object sizes. The HAR model becomes complex under environmental variations or occlusions, as it is based on fixed spatial and temporal features that may not adapt well to dynamic or cluttered backgrounds. The combination of MobileNetV2 for spatial feature extraction and a multilayer GRU for temporal modelling enables the model to capture both spatial details and temporal dependencies.

4. CONCLUSION

This research proposes an SKN-2D-CNN model with ArcFace loss for the effective capture of multi-scale spatial features, leveraging its highly discriminative ability for HAR. The key findings show that integrating the selective kernel mechanism enables the network to dynamically adjust the receptive field, focusing on the most informative spatial patterns and minimizing the impact of irrelevant frames. Moreover, the ArcFace loss function employs an angular margin penalty that enhances intra-class compactness and inter-class separability, resulting in superior performance. The practical implication of this research lies in its ability to enhance the reliability and robustness of video-based HAR, which is significant for surveillance applications. Compared to existing methods like CNN-BiGRU, the proposed SKN-2D-CNN with ArcFace loss achieves superior accuracies of 99.16 and 98.75% on the UFC101 and HMDB-51 datasets, respectively. In the future, improved optimization-based methods will be used to select the most appropriate features, further enhancing model performance in HAR.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Banushri Srinivasaiah	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Jagadeesha Ramegowda		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available in the HMDB-51 at https://www.kaggle.com/datasets/easonlll/HMDB-51 and UFC101 datasets at https://www.kaggle.com/datasets/matthewjansen/ucf101-action-recognition.

REFERENCES

[1] M. A. Khan *et al.*, "Human action recognition using fusion of multiview and deep features: an application to video surveillance," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14885–14911, 2024, doi: 10.1007/s11042-020-08806-9.

[2] A. Hussain, S. U. Khan, N. Khan, M. Shabaz, and S. W. Baik, "AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems," *Engineering Applications of Artificial Intelligence*, vol. 127, 2024, doi: 10.1016/j.engappai.2023.107218.

[3] Y. Kaya and E. K. Topuz, "Human activity recognition from multiple sensors data using deep CNNs," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 10815–10838, 2023, doi: 10.1007/s11042-023-15830-y.

[4] M. S. Raj, S. N. George, and K. Raja, "Leveraging spatio-temporal features using graph neural networks for human activity recognition," *Pattern Recognition*, vol. 150, 2024, doi: 10.1016/j.patcog.2024.110301.

[5] Y. Zhou, J. Xie, X. Zhang, W. Wu, and S. Kwong, "Energy-efficient and interpretable multisensor human activity recognition via deep fused lasso net," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 5, pp. 3576–3588, 2024, doi: 10.1109/TETCI.2024.3430008.

[6] P. Lalwani and G. Ramasamy, "Human activity recognition using a multi-branched CNN-BiLSTM-BiGRU model," *Applied Soft Computing*, vol. 154, 2024, doi: 10.1016/j.asoc.2024.111344.

[7] G. Pareek, S. Nigam, and R. Singh, "Modeling transformer architecture with attention layer for human activity recognition," *Neural Computing and Applications*, vol. 36, no. 10, pp. 5515–5528, 2024, doi: 10.1007/s00521-023-09362-7.

[8] M. Ezzeldin, A. S. Ghoneim, L. Abdelhamid, and A. Atia, "Multi-modal hybrid hierarchical classification approach with transformers to enhance complex human activity recognition," *Signal, Image and Video Processing*, vol. 18, pp. 9375–9385, 2024, doi: 10.1007/s11760-024-03552-z.

[9] H. Park, G. H. Lee, J. Han, and J. K. Choi, "Multiclass autoencoder-based active learning for sensor-based human activity recognition," *Future Generation Computer Systems*, vol. 151, pp. 71–84, 2024, doi: 10.1016/j.future.2023.09.029.

[10] Z. Yang, K. Li, and Z. Huang, "MFCANN: a feature diversification framework based on local and global attention for human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 133, Jul. 2024, doi: 10.1016/j.engappai.2024.108110.

[11] Y. C. Lai, Y. C. Kan, K. C. Hsu, and H. C. Lin, "Multiple inputs modeling of hybrid convolutional neural networks for human activity recognition," *Biomedical Signal Processing and Control*, vol. 92, 2024, doi: 10.1016/j.bspc.2024.106034.

[12] M. A. Al-qaness, A. Dahou, M. A. Elaziz, and A. M. Helmi, "Human activity recognition and fall detection using convolutional neural network and transformer-based architecture," *Biomedical Signal Processing and Control*, vol. 95, 2024, doi: 10.1016/j.bspc.2024.106412.

[13] A. Boudjema, F. Titouna, and C. Titouna, "AReNet: cascade learning of multibranch convolutional neural networks for human activity recognition," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 51099–51128, Nov. 2023, doi: 10.1007/s11042-023-17496-y.

[14] Q. Lin, X. Li, K. Cai, M. Prakash, and D. Paulraj, "Secure internet of medical things (IoMT) based on ECMQV-MAC authentication protocol and EKMC-SCP blockchain networking," *Information Sciences*, vol. 654, 2024, doi: 10.1016/j.ins.2023.119783.

[15] T. Teng, J. Wan, and X. Zhang, "GANCDE: neural networks based on graphs and attention neural control differential equations for human activity recognition," *Knowledge and Information Systems*, vol. 66, pp. 6213–6240, Jun. 2024, doi: 10.1007/s10115-024-02154-y.

[16] T. Ahmad, J. Wu, H. S. Alwageed, F. Khan, J. Khan, and Y. Lee, "Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection," *IEEE Access*, vol. 11, pp. 33148–33159, 2023, doi: 10.1109/ACCESS.2023.3263155.

[17] K. P. Sinha and P. Kumar, "Human activity recognition from UAV videos using an optimized hybrid deep learning model," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 51669–51698, 2024, doi: 10.1007/s11042-023-17289-3.

[18] A. Kushwaha, A. Khare, and O. Prakash, "Micro-network-based deep convolutional neural network for human activity recognition from realistic and multi-view visual data," *Neural Computing and Applications*, vol. 35, no. 18, pp. 13321–13341, 2023, doi: 10.1007/s00521-023-08440-0.

[19] N. Varshney and B. Bakariya, "Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 42117–42129, 2022, doi: 10.1007/s11042-021-11220-4.
[20] T. Ahmad and J. Wu, "SDIGRU: spatial and deep features integration using multilayer gated recurrent unit for human activity recognition," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 973–985, 2023, doi: 10.1109/TCSS.2023.3249152.
[21] Easonlll and SCNUlyx, "HMDB-51," *Kaggle*. Accessed: Jan. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/easonlll/hmdb51
[22] M. Jansen, "UCF101 - action recognition," *Kaggle*. Accessed: Jan. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/matthewjansen/ucf101-action-recognition
[23] P. K. Sahoo *et al.*, "An improved VGG-19 network induced enhanced feature pooling for precise moving object detection in complex video scenes," *IEEE Access*, vol. 12, pp. 45847–45864, 2024, doi: 10.1109/ACCESS.2024.3381612.
[24] A. Karacı, "VGGCOV19-NET: automatic detection of COVID-19 cases from X-ray images using modified VGG19 CNN architecture and YOLO algorithm," *Neural Computing and Applications*, vol. 34, no. 10, pp. 8253–8274, 2022, doi: 10.1007/s00521-022-06918-x.
[25] Y. Ji, D. Wang, Q. Li, T. Liu, and Y. Bai, "Global wildfire danger predictions based on deep learning taking into account static and dynamic variables," *Forests*, vol. 15, no. 1, Jan. 2024, doi: 10.3390/f15010216.

BIOGRAPHIES OF AUTHORS

Banushri Srinivasaiah is a research scholar pursuing the Ph.D. degree in Computer Science and Engineering. She is currently working as an assistant professor, with 8+ years of experience in the computer science field with a focus on deep learning, with the Visveswaraya Institute of Technology (VTU), Belagavi, Karnataka, India under the supervision of Dr. Jagadeesha R. Her research interests include human activity recognition, sequence learning for visual data, machine learning, deep learning, context-based image indexing and retrieval, facial expression prediction, and computer vision. She can be contacted at email: banushri914@gmail.com.

Jagadeesha Ramegowda is currently working as professor and head of research center in Department of CSE at Impact College of Engineering Applied Sciences, Bengaluru. Ph.D. degree in Computer Science and Engineering from VTU, Belagavi, Karnataka, India. He has more than 16 years of experience in teaching and research. His research interests include artificial intelligence, computer networks, cyber security, data mining, and cloud computing. He can be contacted at email: jagdish.mtech@gmail.com.