Enhancing Fitness Movement Recognition with Attention Mechanism and Pre-Trained Feature Extractors

Shanjid Hasan Nishat*, Srabonti Deb†, Mohiuddin Ahmed‡
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh
Emails: *shanjidhasan202@gmail.com, †srabontideb23@gmail.com, ‡mohiuddin.nirob.mn@gmail.com

Abstract—Fitness movement recognition, a focused subdomain of human activity recognition (HAR), plays a vital role in health monitoring, rehabilitation, and personalized fitness training by enabling automated exercise classification from video data. However, many existing deep learning approaches rely on computationally intensive 3D models, limiting their feasibility in real-time or resource-constrained settings. In this paper, we present a lightweight and effective framework that integrates pre-trained 2D Convolutional Neural Networks (CNNs) such as ResNet50, EfficientNet, and Vision Transformers (ViT) with a Long Short-Term Memory (LSTM) network enhanced by spatial attention. These models efficiently extract spatial features while the LSTM captures temporal dependencies, and the attention mechanism emphasizes informative segments. We evaluate the framework on a curated subset of the UCF101 dataset, achieving a peak accuracy of 93.34% with the ResNet50-based configuration. Comparative results demonstrate the superiority of our approach over several state-of-the-art HAR systems. The proposed method offers a scalable and real-time-capable solution for fitness activity recognition with broader applications in vision-based health and activity monitoring.

Index Terms—Fitness Movement Recognition, ResNet50, EfficientNet, Vision Transformer, Long Short-Term Memory, Attention Mechanism, Transfer Learning, UCF101

I. INTRODUCTION

Human activity recognition (HAR), a fundamental aspect of intelligent vision systems, enables computational interpretation of physical activities from video or sensor data. HAR continues to play a pivotal role in various domains including video surveillance [1], healthcare [2], human-computer interaction [3], and autonomous driving [4]. A prominent subfield within this domain is fitness movement recognition, which focuses specifically on identifying exercise-related actions for applications such as personalized monitoring, real-time feedback, and rehabilitation support [5]. Accurate recognition of fitness activities contributes to improved health management and disease prevention [6]. Therapists emphasize the importance of activity-based interventions, and modern technological platforms such as wearable devices and tele-exercise systems facilitate these goals by offering real-time feedback and sustained user engagement [7].

The progression of fitness technology reflects the growing shift toward data-driven solutions for promoting active lifestyles. From early pedometers to smartwatches and tele-exercise platforms, modern systems now offer features such as step counting, heart rate monitoring, and personalized exercise tracking [8]. Studies highlight their effectiveness in boosting short-term physical activity and reducing BMI, though sustained behavioral change remains underexplored [9]. Historically evolving from mechanical equipment to digital gym machines with metric displays [10], today's technologies leverage artificial intelligence and large-scale data to deliver adaptive and user-centric fitness insights.

While the importance of fitness movement recognition is widely acknowledged, existing computational models still face challenges. Traditional approaches relying on handcrafted features were limited in accuracy and scalability [11]. More recently, 3D Convolutional Neural Networks (CNNs) [12] have shown strong performance in activity recognition but suffer from high computational costs, which restrict their applicability in real-time and resource-constrained environments such as smartphones and wearable devices [13]. This limitation highlights the need for lightweight yet accurate frameworks capable of balancing performance with efficiency. To address these challenges, this research introduces a novel architecture that integrates pre-trained feature extractors with Long Short-Term Memory (LSTM) [14] networks enhanced by an attention mechanism. Specifically, pre-trained CNNs such as ResNet [15], EfficientNet [16], and Vision Transformers (ViT) [17], trained on large-scale datasets like ImageNet, are used to extract discriminative spatial features from fitness video frames. These features are then processed by an LSTM network to capture temporal dynamics, while the attention mechanism further improves performance by adaptively focusing on the most informative frames or features during recognition. The major contributions of this work include-

- Development of a lightweight hybrid framework integrating 2D CNN, LSTM, and attention mechanism for efficient fitness movement recognition, with significantly fewer parameters compared to 3D CNN-based models.
- Comparative evaluation of multiple deep learning models to identify the most accurate and computationally efficient approach for fitness movement recognition.

- Advancement of fitness monitoring technologies by providing a practical recognition system suitable for healthcare, personalized coaching, and rehabilitation.

The rest of the paper is structured as follows: Section II reviews key techniques and models in movement recognition. Section III outlines the proposed architecture of the model, dataset description, data preprocessing, and a detailed description of model components. Section IV discusses experimental setup, hyperparameters, and results. Finally, the last section provides the study's final analysis and prospective areas for future exploration.

II. RELATED WORKS

The field of HAR has progressed notably with the advent of deep learning, shifting from conventional handcrafted methods toward models that automatically extract spatial and temporal representations from video and sensor data. Early efforts such as those by Chen et al. [18] introduced supervised hashing frameworks for efficient video retrieval, emphasizing scalability through compact binary encodings. While such approaches were suitable for large-scale search tasks, they lacked the temporal modeling necessary for fine-grained motion recognition, especially in the fitness domain.

Vrskova et al. [19] proposed a lightweight 3D CNN that balances accuracy and computational cost. Despite its efficiency, the architecture was limited in capturing subtle motion patterns and overlapping actions. To overcome the challenges of generalization and temporal awareness, Zhou et al. [20] introduced a knowledge distillation method based on Vision Transformers, transferring spatial attention cues from a large supervisory model to a smaller one. Although this reduced model size, transformer-based frameworks typically require heavy pretraining and remain resource-intensive for real-time applications.

Huang et al. [21] addressed class scalability through a zero-shot action recognition model using a dual-GAN to synthesize visual features of unseen classes from semantic embeddings. Yet, its performance declined on repetitive and temporally dependent motions typical in fitness contexts.

Focusing on explainability and privacy, Bsoul [22] proposed a graph-based classifier leveraging skeletal keypoints. This offered interpretable pose-based predictions but was vulnerable to occlusions and noisy environments, limiting robustness. Dasan et al. [23] accelerated mobile HAR using MoViNet with reduced latency and model size, though the simplified design hindered the discrimination of visually similar exercises.

Maji et al. [24] tackled low-latency activity recognition via a hybrid model combining convolutional layers with attention-augmented GRUs using wearable sensor input. Despite its temporal precision, reliance on sensor data constrains its adaptability for vision-only environments better suited for fitness tracking using camera feeds.

In light of these developments, our proposed model adopts a balanced hybrid architecture, integrating pre-trained 2D CNNs for spatial representation with LSTM layers for temporal dynamics, further enhanced by a spatial attention mechanism.

Unlike zero-shot or sensor-based methods, our pipeline directly processes visual inputs, making it broadly applicable and independent of external sensing hardware. Moreover, it avoids the overhead of transformer-heavy models while retaining the benefit of attention-based focus on discriminative temporal segments. This design enables effective recognition of fine-grained fitness activities in real time, with improved generalization and scalability across environments.

III. MATERIALS AND METHODS

Our proposed architecture for fitness movement recognition is illustrated in Fig.1. The pipeline is designed to effectively capture both spatial and temporal dynamics from video data, ensuring robust classification of exercise activities.

At the initial stage, a 2D feature extractor processed each frame of the input video sequence, capturing spatial patterns. These frame-wise features were passed through a TimeDistributed layer to maintain consistency across all frames. The extracted features were then flattened and fed into an LSTM layer, which models temporal dependencies and sequential dynamics. To enhance focus on the most informative regions, a spatial attention mechanism was applied to the LSTM output. The resulting attention weights were combined with the temporal features, emphasizing relevant spatial cues while suppressing irrelevant ones. The attended features were aggregated using global average pooling across the temporal dimension, producing a compact and discriminative representation. Finally, this representation was passed through a dense layer with softmax activation, enabling classification into the predefined fitness movement categories.

A. Dataset Description

In this study, we used the UCF101 dataset [25] to evaluate the performance of the proposed fitness movement recognition framework. UCF101 (University of Central Florida 101) is a widely adopted action recognition dataset consisting of 13,320 video clips categorized into 101 human activity classes. These classes span various domains, including sports, daily activities, and fitness-related exercises, making it suitable for human activity recognition research.

Each class in the dataset contains approximately 133 video clips on average, offering sufficient diversity and volume for training and evaluation. For this work, we selected a subset of fitness-related classes that align with the objectives of fitness movement recognition.

B. Dataset Preprocessing

To prepare the video data for model training, a standardized preprocessing pipeline was applied to ensure consistency across samples and compatibility with the deep learning architecture. The key steps include:

1) *Frame Extraction:* From each video, 20 frames were uniformly sampled at equal intervals to capture temporal dynamics. Given a video with N frames, the selected frames $\{f_{t_1}, f_{t_2}, \ldots, f_{t_{20}}\}$ were indexed using:

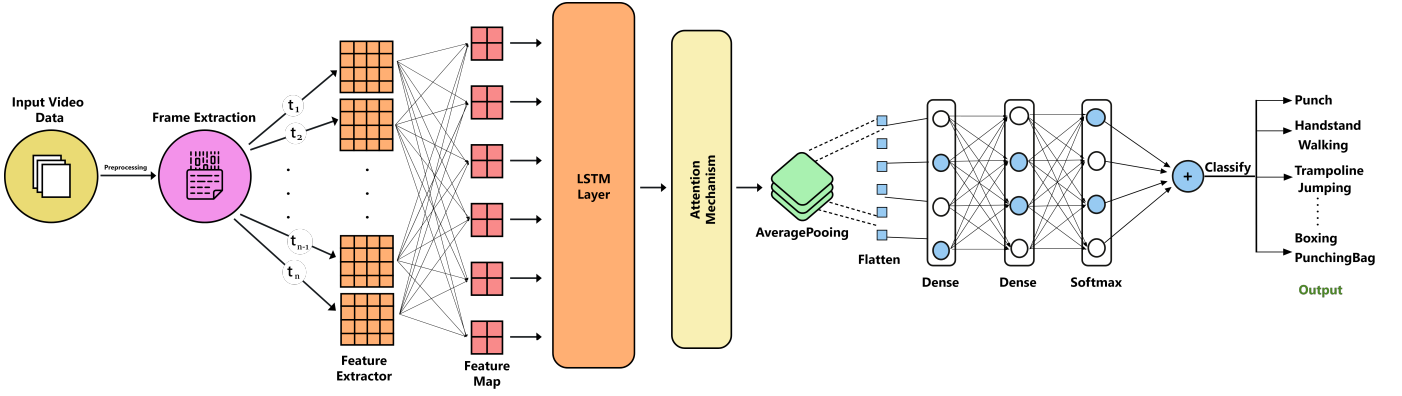$$t_i = \left\lfloor \frac{i \times N}{21} \right\rfloor \quad (1)$$

Fig. 1: Architecture of Proposed Methodology

This sampling strategy maintains temporal consistency while reducing computational load.

2) *Frame Resizing:* All frames were initially resized to 64×64 for storage and processing efficiency. Before feature extraction using pre-trained models (trained on ImageNet with 224×224 inputs), frames were upscaled to 224×224 to ensure compatibility.

3) *Normalization:* Pixel values were normalized to the range [0, 1] by dividing each pixel intensity by 255:

$$\text{Normalized Value} = \frac{\text{Original Value}}{255} \quad (2)$$

These preprocessing steps ensured a uniform and compact input representation, facilitating efficient training and stable convergence of the model.

C. 2D Feature Extractor

To extract spatial features from individual video frames, a convolutional layer was employed as the foundational component. This layer applies a set of learnable filters (kernels) across the input image to detect local patterns such as edges, textures, and shapes. For a single-channel input image X and a filter W, the convolution operation is expressed as:

$$F = X * W + b \quad (3)$$

where $*$ denotes the convolution operator, F is the resulting feature map, and b is a bias term.

Several core mechanisms are integrated into the convolutional process:

- *Feature Mapping:* Convolution produces feature maps that highlight the activation of specific filters at different spatial locations.
- *Weight Sharing:* The same filter is applied across all regions of the input, significantly reducing the number of parameters and enhancing model generalization.
- *Nonlinear Activation:* A ReLU activation function is applied to the output of the convolution to introduce nonlinearity.
- *Padding and Stride:* Padding preserves spatial dimensions at the borders, while stride controls the filter's step size during scanning.

D. Pre-trained Vision Transformer (ViT)

The Vision Transformer (ViT) is a transformer-based architecture originally developed for natural language processing and later adapted to vision tasks by representing images as sequences of patches.

1) *Patch Embedding:* The input image X is divided into N non-overlapping patches, each of which is linearly projected into a d-dimensional embedding space:

$$E = \text{Linear}(\text{Patchify}(X)) \quad (4)$$

where $E \in \mathbb{R}^{N \times d}$ denotes the patch embedding matrix.

2) *Positional Encoding:* To compensate for the lack of spatial locality in transformers, positional encodings P are added to the patch embeddings to preserve spatial order.

3) *Transformer Encoder:* The sequence of embedded patches is processed by a stack of L transformer encoder blocks. Each block consists of multi-head self-attention and feedforward layers. The output of the l-th encoder block is defined as:

$$H^{(l)} = \text{TransformerEncoderBlock}(H^{(l-1)}), \quad (5)$$

where

$$H^{(0)} = E + P \quad (6)$$

4) *Classification Head:* The final transformer output $H^{(L)}$ is passed through a linear layer followed by softmax for classification:

$$Y = \text{Softmax}(\text{Linear}(H^{(L)})) \quad (7)$$

By pre-training on large-scale datasets like ImageNet, ViT learns rich visual representations. Replacing the classification head with a task-specific layer allows the model to be fine-tuned for fitness movement recognition.

E. Pre-trained EfficientNet

EfficientNet is a CNN architecture designed to balance accuracy and efficiency through a compound scaling strategy that uniformly scales model depth, width, and input resolution. The pre-trained EfficientNet model, trained on ImageNet,

serves as a lightweight feature extractor for various vision tasks.

1) *Efficient Building Blocks:* The architecture is composed of optimized modules such as depthwise separable convolutions, inverted residual blocks, and squeeze-and-excitation mechanisms, all contributing to reduced computational costs without sacrificing performance.

2) *Transfer Learning:* After pre-training on ImageNet, the classification layers are removed, and the model is fine-tuned with task-specific heads for applications like image classification or movement recognition. This enables EfficientNet to generalize well even on smaller, domain-specific datasets.

F. Pre-trained ResNet50

ResNet50 is a deep CNN consisting of 50 layers which is based on the Residual Network (ResNet) architecture. It introduces skip connections to mitigate the vanishing gradient problem and enable effective training of very deep networks.

1) *Residual Blocks:* Each residual block contains convolutional layers with a shortcut connection that bypasses intermediate layers, allowing gradients to propagate more easily and improving convergence.

2) *Bottleneck Design:* ResNet50 employs a bottleneck structure within its residual blocks, consisting of a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution. This design reduces computational cost while preserving model capacity.

3) *Transfer Learning:* Pre-trained on the ImageNet dataset, ResNet50 can be fine-tuned for downstream tasks by replacing its classification head with task-specific layers. This approach enables the reuse of learned visual features on smaller or domain-specific datasets.
ResNet50's depth and skip connections make it a reliable feature extractor for fitness movement recognition.

G. TimeDistributed Layer

The TimeDistributed layer applies the same operation independently to each time step in a sequence, making it ideal for tasks like processing video frames. Formally, for an input sequence $X = \{X_1, X_2, \ldots, X_T\}$, the output is computed as:

$$Y_t = f(X_t) \tag{8}$$

where f is the operation applied at each time step t. This layer facilitates efficient parallel computation and is commonly used to connect CNNs to RNNs in video or time-series applications.

H. LSTM Layer

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that effectively captures long-term dependencies in sequential data. It is particularly well-suited for tasks that require modeling temporal dynamics, such as video-based human activity recognition.

An LSTM unit consists of a memory cell that retains information over time and three essential gating mechanisms : the input gate, forget gate, and output gate which regulate the flow of information into, within, and out of the cell. These gates enable selective memory updates and prevent vanishing gradients during training.

In our framework, the LSTM layer processes sequential frame-wise features extracted from video inputs. Its ability to retain relevant historical context across time steps significantly contributes to modeling the temporal structure of fitness movements, leading to improved classification accuracy.

I. Spatial Attention Mechanism

The temporal attention mechanism enables the network to focus on informative regions of the input while suppressing irrelevant ones.

Given an input feature map $H \in \mathbb{R}^{T \times d}$ from the LSTM layer, attention weights $A \in \mathbb{R}^T$ are computed as:

$$A = \text{Softmax}(W_a \cdot \text{ReLU}(W_h H + b_h) + b_a) \tag{9}$$

where W_h, W_a are learnable parameters and b_h, b_a are bias terms.

- *Attention Scores:* Assign importance values to each time step in the sequence.
- *Feature Weighting:* Weighted features emphasize relevant spatial cues while downplaying less informative ones.
- *Performance Gain:* By highlighting critical regions, temporal attention improves accuracy and robustness in sequence-based recognition.

J. Dense Layer for Classification

The Dense (fully connected) layer maps the extracted feature representation to class predictions through a linear transformation followed by an activation function. For an input vector $X \in \mathbb{R}^n$, the output is computed as:

$$Z = W \cdot X + b \tag{10}$$

where $W \in \mathbb{R}^{m \times n}$ is the weight matrix and $b \in \mathbb{R}^m$ is the bias.

A softmax activation is applied to Z to generate class probabilities:

$$P_i = \frac{e^{Z_i}}{\sum_{j=1}^{C} e^{Z_j}} \tag{11}$$

where C is the number of classes. The class with the highest probability is selected as the prediction.

As the final layer of the model, the Dense layer enables supervised learning by mapping high-level features to predefined fitness movement categories.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup and Hyperparameters

All experiments were conducted on a system equipped with an Intel Core i9 11th Gen processor, 32GB RAM, and an NVIDIA RTX A4500 GPU with 36GB of dedicated memory. The proposed model was trained and evaluated on a curated subset of the UCF101 dataset, comprising 35 fitness-related activity classes. For each class, 80% of the videos were initially allocated to the training and validation set, and the

remaining 20% were reserved for testing. Within the training and validation set, 80% was used for training and 20% for validation.

Training was performed using the Adam optimizer with a learning rate of 0.0001, a batch size of 32, and categorical cross-entropy loss. Early stopping was employed to prevent overfitting based on validation loss monitoring.

To compare model efficiency, Table I presents the total, trainable, and non-trainable parameters of each variant.

TABLE I: Parameter Counts of Models

Model	Total Parameters	Trainable	Non-Trainable
ViT+LSTM	87 636 195	215 523	87 420 672
EfficientNet+LSTM	9 595 590	9 553 567	42 023
ResNet50+LSTM	**27 852 579**	**27 799 459**	**53 120**
ConvLSTM (Baseline)	133 835	133 835	0

B. Classification Results

To evaluate the effectiveness of various spatial-temporal modeling strategies, four deep learning models were implemented. Each model varied in its choice of feature extractor while keeping the temporal modeling and attention components consistent.

Model 1: LRCN (Conv + LSTM + Attention): The baseline LRCN model achieved 73.82% accuracy and 91.12% top-5 accuracy. Despite stable training, Fig. 2 shows that its recall and F1 scores were lower due to misclassification of visually similar actions.
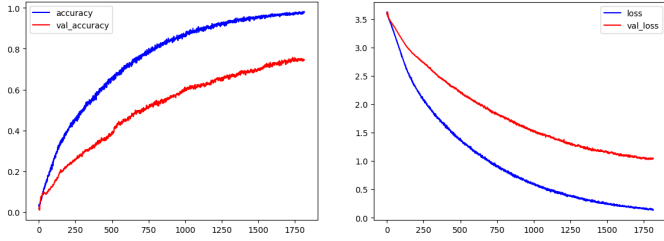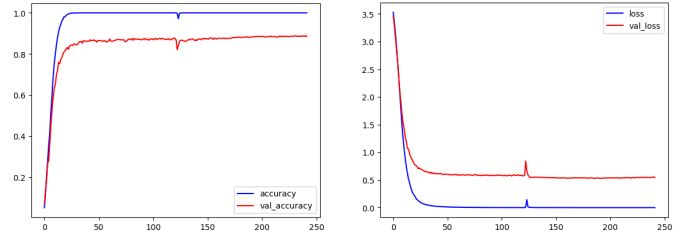

(a) Accuracy vs. Validation Accuracy (LRCN + Attention)   (b) Loss vs. Validation Loss (LRCN + Attention)

Fig. 2: Training Performance Curves of LRCN + Attention Model

Model 2: ViT + LSTM: As shown in Fig. 3, incorporating a Vision Transformer significantly improved performance to 88.61% accuracy. However, the model required higher memory and longer training time due to the transformer complexity.
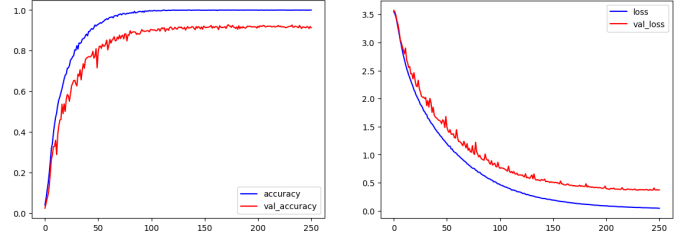
Model 3: EfficientNet + LSTM + Attention: The EfficientNet-based model achieved an accuracy of 92.01%. The corresponding training and validation performance trends, including loss and accuracy metrics over epochs, are depicted in Fig. 4.

Model 4: ResNet50 + LSTM + Attention: Our best-performing model combined ResNet50 with LSTM and attention, achieving 93.34% accuracy and 97.04% top-5 accuracy. As illustrated in Fig. 5, the model maintained stable learning

(a) Accuracy vs. Validation Accuracy (ViT)   (b) Loss vs. Validation Loss (ViT)

Fig. 3: Training Performance Curves of ViT + LSTM Model

(a) Accuracy vs. Validation Accuracy (EfficientNet)   (b) Loss vs. Validation Loss (EfficientNet)

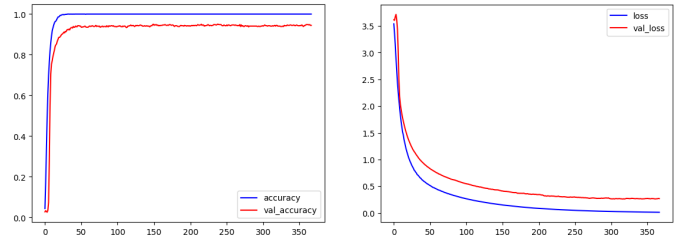Fig. 4: Training Performance Curves of EfficientNet + LSTM + Attention Model

with strong generalization, aided by ResNet's residual connections and temporal refinement via attention.

Table II presents the test set performance comparison across all four proposed models, detailing accuracy, precision, recall, F1 score, and top-5 accuracy.

TABLE II: Comparative Test Performance of Models

Model	Acc.	Top-5 Acc.	Prec.	Recall	F1
LRCN	73.82%	91.12%	75.49%	73.82%	73.57%
ViT+LSTM	88.61%	95.71%	89.22%	88.61%	88.53%
EffNet+Attn	92.01%	97.04%	92.37%	92.01%	91.88%
ResNet50+Attn	**93.34%**	**97.04%**	**93.59%**	**93.34%**	**93.23%**

Finally, Table III compares our top-performing model against recent state-of-the-art HAR frameworks from existing literature. It demonstrates that our ResNet50-based model

(a) Accuracy vs. Validation Accuracy (ResNet50)   (b) Loss vs. Validation Loss (ResNet50)

Fig. 5: Training Performance Curves of ResNet50 + LSTM + Attention Model

achieves competitive or superior accuracy, validating its effectiveness in recognizing fitness activities.

TABLE III: Performance Comparison with Existing HAR Approaches

Work	Model	Accuracy (%)
Chen et al. [18]	Supervised Hashing (Vision)	85.60
Vrskova et al. [19]	Lightweight 3D CNN (Vision)	89.50
Zhou et al. [20]	ViT + Distillation	92.90
Huang et al. [21]	Dual-GAN (Zero-Shot, Vision)	89.82
Bsoul [22]	Graph-Based Skeleton Classifier	91.80
Dasan et al. [23]	MoViNet-A0 (Mobile Vision)	91.20
Maji et al. [24]	GRU + Attention (Sensor)	93.20
Ours	**ResNet50 + LSTM + Attention**	**93.34**

V. CONCLUSION AND FUTURE WORK

This paper presents a lightweight framework for fitness movement recognition using video data, combining pre-trained 2D CNNs with LSTM and attention mechanisms to capture both spatial and temporal dynamics. Among the tested configurations, the ResNet50-based architecture with LSTM and attention mechanism achieved the highest accuracy, demonstrating its strength in recognizing fine-grained and repetitive exercise actions. The proposed system contributes a practical solution for real-time activity recognition, with strong applicability in fitness tracking, healthcare, rehabilitation, and coaching. Despite its promising performance, several limitations remain. The model's accuracy is highly influenced by the diversity and balance of the training dataset, and deep neural networks still pose challenges in terms of computational cost and interpretability.

In our future research, we will focus on expanding the dataset for broader generalization, optimizing the architecture for deployment on resource-constrained devices, and exploring multimodal fusion with wearable sensors. Enhancing user interaction through real-time, explainable feedback is also a key direction to improve accessibility and effectiveness in real-world fitness applications.

REFERENCES

[1] M. Khan *et al.*, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimedia Tools and Applications*, vol. 83, pp. 5759–5783, 2020.

[2] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, Oct. 2014.

[3] R. Cheng, A. L. Yuille, and E. G. Learned-Miller, "Human activity recognition from RGB-D data with deep learning," *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–36, Sep. 2021.

[4] A. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB depth sensors for fall detection: A review," *IEEE Sensors Journal*, vol. 17, no. 12, pp. 3585–3604, 2017.

[5] Y. Liu *et al.*, "Vision-based human activity recognition: A survey of recent developments," *Sensors*, vol. 20, no. 4, pp. 1–23, Feb. 2020.

[6] C. D. Kidd and R. W. Picard, "Evaluation of affective response to robotic movement," *Proceedings of the 2005 IEEE-RAS International Conference on Humanoid Robots*, pp. 380–385, Dec. 2005.

[7] Y. Chen *et al.*, "Smart wearable devices in cardiovascular care: A systematic review and meta-analysis of diagnostic and therapeutic effects," *European Heart Journal – Digital Health*, vol. 1, no. 1, pp. 21–29, 2020.

[8] A. S. M. Shahjalal, A. K. M. Mahbubur Rahman, and S. R. Rizvi, "Smart Fitness Tracking System for Monitoring and Recommending Exercise Using Machine Learning," *IEEE Access*, vol. 10, pp. 66301–66315, 2022.

[9] A. Nayak, P. A. Nayak, P. K. Mohapatra, A. Das, and S. Kumar, "IoT-enabled smart fitness: A system for providing feedback and motivation to users," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 2293–2308, 2021.

[10] K. Hemalatha, A. Suresh, and P. Vetrivelan, "Smart Gym System Based on IoT for Health and Fitness Monitoring," *International Journal of Engineering Research and Applications*, vol. 10, no. 3, pp. 32–36, 2020.

[11] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 1–43, Apr. 2011.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021.

[18] L. Chen, Y. Zhang, X. Jin, and L. Zhang, "Supervised hashing with structural balance for scalable video retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 3067–3079, 2021.

[19] R. Vrskova, M. Hradiš, and P. Zemcík, "Efficient lightweight 3D CNN for human action recognition," in *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, 2022, pp. 281–289.

[20] K. Zhou, T. Zhang, X. Cui, M. Yang, and C. Ma, "Knowledge distillation from vision transformer to convolutional neural networks for action recognition," *Neurocomputing*, vol. 503, pp. 228–237, 2022.

[21] Z. Huang, J. Chen, and C. Zhang, "Zero-shot action recognition with dual adversarial generative networks," *IEEE Transactions on Image Processing*, vol. 32, pp. 1283–1296, 2023.

[22] M. Bsoul, "Pose-aware human activity recognition using graph convolutional networks for privacy-preserving environments," in *Proc. ACM International Conference on Multimedia*, 2025.

[23] P. Dasan, R. Chhabra, and M. Gupta, "Efficient mobile action recognition using MoViNet and attention pooling," in *Proc. IEEE International Conference on Pervasive Computing*, 2025.

[24] S. Maji, R. Roy, and P. Dutta, "A hybrid model for human activity recognition using attention-based GRU and CNN on wearable sensor data," *Procedia Computer Science*, vol. 219, pp. 317–324, 2023.

[25] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," University of Central Florida, CRCV-TR-12-01, Nov. 2012.