International Journal of Advances in Applied Sciences (IJAAS)
Vol. 15, No. 2, June 2026, pp. 437~450
ISSN: 2252-8814, DOI: 10.11591/ijaas.v15.i2.pp437-450

Real-time human activity recognition using deep learning techniques for the next-generation healthcare system

Subrata Kumer Paul[1,2], Rakhi Rani Paul[1,2], Md. Ekramul Hamid[1], Md. Rafiqul Islam (Rafiq)[2]
[1]Department of Computer Science and Engineering, Faculty of Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh
[2]Department of Computer Science and Engineering, Faculty of Electrical and Computer Engineering,
Bangladesh Army University of Engineering and Technology (BAUET), Natore-6431, Bangladesh

Article Info

Article history:

Received Sep 6, 2024
Revised Mar 6, 2026
Accepted Apr 22, 2026

Keywords:

Convolutional long short-term memory
EfficientNetB0
EfficientNetB0ConvLSTM
HMDB51 dataset
Healthcare
Human activity recognition

ABSTRACT

In today's world, healthcare systems are being built with human activity recognition (HAR) to help the elderly, disabled, and children's activities by constantly observing their behavior. However, HAR using computer vision and traditional machine learning techniques is not an efficient use of healthcare system resources because of potential and accuracy issues. This study aims to examine the use of deep learning techniques in real-time HAR. This proposal is a hybrid method that utilizes the EfficientNetB0 architecture and a convolutional extension of a long short-term memory network (EfficientNetB0ConvLSTM), to achieve human-like intelligence. The EfficientNetB0 is utilized to extract image features, and convolutional long short-term memory (ConvLSTM) is utilized to categorize six human actions to recognize human activities. This approach leverages the strengths of convolutional neural networks (CNNs) in extracting spatial features from video frames and LSTMs in capturing temporal dependencies within activity sequences. Firstly, an extensive investigation is conducted on existing literature studies to select a suitable dataset. Next, the proposed method was evaluated on the challenging HMDB51 video datasets and finally achieved an accuracy of 89.22%, which is significantly higher than other methods on this dataset. This outcome shows the potential of EfficientNetB0ConvLSTM for real-time HAR applications like healthcare.

Corresponding Author:

Md. Ekramul Hamid
Department of Computer Science and Engineering, Faculty of Engineering, University of Rajshahi
Rajshahi-6205, Bangladesh
Email: ekram_hamid@ru.ac.bd

1. INTRODUCTION

Human activity recognition (HAR) is a critical area in the healthcare sector, as it enables the monitoring of patients, supports elderly care, and facilitates rehabilitation [1]. Traditional HAR systems often face challenges, including latency and accuracy issues, which hinder their effectiveness in real-time applications. These systems are designed to automatically recognize physical activities by analyzing data collected from sensors or video footage. However, existing approaches often face challenges in simultaneously achieving high accuracy and low latency, both of which are critical for applications such as continuous monitoring, timely intervention, and personalized healthcare [2]. A range of methods have been explored to improve HAR performance, from traditional machine learning algorithms to advanced deep learning models such as convolutional neural networks (CNNs) [3] and recurrent neural networks (RNNs) [4]. Current HAR systems face several significant limitations that affect their performance [5]. Achieving high accuracy is one of the challenges, particularly in dynamic real-world conditions. The ultimate goal is to

create a system capable of accurately recognizing human activities in real-time, which can be effectively integrated into healthcare settings for continuous monitoring, timely interventions, and improved patient outcomes. The study emphasizes the importance of the HAR dataset and enhancing human-computer interaction (HCI), particularly in healthcare sectors. Our analysis focuses on the HMDB51 video dataset, selecting six key medical activities such as fall_floor, walk, stand, eat, sit, and drink due to their importance in healthcare. The proposed EfficientNetB0 architecture and a convolutional extension of a long short-term memory network (EfficientNetB0ConvLSTM) model integrates EfficientNetB0 for spatial feature extraction and convolutional long short-term memory (ConvLSTM) for capturing temporal dependencies within activity sequences. The model processes red, green, blue (RGB) video frames through the spatial stream, while temporally stacked optical flow frames are processed through the temporal stream. For training, RGB video frames are resized to 224×224 pixels to align with the model's input requirements. The mini-batch size is fixed at 16 due to hardware limitations. Both streams are trained with a learning rate of 10^{-2}, and the Swish activation function is integrated into the EfficientNetB0 architecture. Average pooling is employed in the pooling layers, while the final output layer uses the SoftMax function, optimized with categorical cross-entropy as the loss function. To mitigate the overfitting problem, a dropout rate of 0.20 is applied in this research after the experiment.

The paper is structured into seven sections. Section 1 introduces the topic and reviews related work. Section 2 describes the dataset and preprocessing steps. Section 3 outlines the proposed methodology. Section 4 presents the experimental results, while section 5 discusses the real-time implementation. Section 6 highlights the limitations and suggests future research directions. Finally, section 7 concludes the paper, followed by acknowledgments and references.

Table 1 summarizes key contributions in the field of HAR using deep learning approaches, with a particular focus on studies that employ the HMDB51 dataset. Different deep-learning algorithms have been utilized for HAR using the HMDB51 dataset, all of which have different strengths, accuracies, and research gaps. Where CNN, long short-term memory (LSTM) networks, and hybrid models like the combination of the CNN + LSTM are commonly used [6]. This analysis underscores the growing prominence of the HMDB51 dataset in HAR research while also pointing out the broader trends in the field.

Table 1. HAR on the HMDB51 dataset (sorted by publication year)

Citations	Year	Used methods	Accuracy (%)	Research gaps
[7]	2022	Semi-supervised temporal gradient learning	75.91	Need for improved semi-supervised learning methods for action recognition in limited labeled data scenarios.
[8]	2022	Singular value thresholding (SVT) (linear) SVT (fine-tune)	57.82 67.21	Lack of robust self-supervised models for action recognition in video sequences.
[9]	2022	Adapting vision transformers (Adaptformer)	51.03	In particular, it emphasizes the need to improve the scalability and adaptability of vision transformers for diverse real-world visual recognition tasks.
[10]	2023	SVFormer-S SVFormer-B	59.71 68.22	The challenge of achieving high accuracy in action recognition with limited labeled video data.
[11]	2023	Critical discourse analysis (CDA) + CNN, Bidirectional gated recurrent unit (GRU)	79.38	Difficulty in handling complex dynamic activities with traditional deep learning models.
[12]	2024	CNN + CAM + AE	77.29	To improve human action recognition accuracy in low-resolution frames while maintaining efficiency.
[13]	2024	SICEC	88.70	Computational cost high in AI systems
[14]	2024	Bidirectional LSTM (BiLSTM)	76.30	Lack of dynamic HAR
[15]	2024	DMSA-UNet	81.20	Lack of efficient methods for accurate medical image segmentation in complex, low-contrast, and noisy scenarios
[16]	2024	STM	80.40	Need for better action recognition methods combining spatiotemporal and motion encoding with efficiency.

The proposed works present findings, limitations, and proposed solutions that are novel contributions, leading to the development of affordable, automated, and intelligent systems that are accessible to all. Preventing false recognition is a significant challenge in intelligent surveillance systems. By improving the accuracy of human activity detection, particularly in medical contexts, there is a direct reduction in the likelihood of false identifications. To address these challenges, this study makes several key contributions.

Introduce an innovative hybrid algorithm combining EfficientNetB0 and ConvLSTM for intelligent surveillance, and enhance the capabilities of the HAR system. Identify and address specific research gaps in the HMDB51 dataset, particularly concerning low accuracy in different medical classes. Contribute to the

dataset selection and the medical class selection process. Design a deep learning model based on EfficientNetB0ConvLSTM, capable of classifying the selected six human activities. Select the most relevant features from video data to reduce computational complexity and enhance model performance. Develop an innovative hybrid model for a HAR-based recognition system with an average accuracy of 89.22%. Enhance the ability to capture spatial features through CNNs and temporal dependencies through LSTMs with the hybrid ENConvLSTM combination. Compare model performances across different deployment scenarios, providing insights into the adaptability and robustness of the proposed method.

2. DATASET SELECTION AND ITS DESCRIPTION

The main focus of this analysis is on three criteria that are used to select the dataset. The proposed activity recognition system's implementation is made easier by setting these criteria. The availability of the activities listed in Table 2 serves as the primary selection criterion. The second criterion considers the similarity between the selected activity videos and those specific to medical-related classes. The third requirement focuses on the richness of features within the video datasets. Furthermore, this study examines six HAR-related video datasets, which are summarized in Table 2.

Table 2. HAR datasets with representative activity classes

Dataset name	Total samples data	Total classes	List of class activities
ActivityNet [17]	21,313	200	drinking, eating, jumping, running, sitting, walking, and swimming
Charades [18]	66,493	157	drinking, eating, running, sitting, and walking
HMDB51 [19]	6,766	51	fall, floor, walk, stand, eat, sit, and drink
NTU RGB+D [20]	56,880	60	falling, vomiting, neck pain, and yawning
STAIR Actions [21]	1,09,478	100	eating, drinking, reading, sitting, and walking
UCF101 [22]	13,320	101	jumping, kicking, running, eating, and falling

Figure 1 presents the dataset analysis and selection process used in this study. It compares activity overlaps among benchmark HAR datasets and highlights the selected HMDB51 dataset for healthcare-related activity recognition. Figure 1(a) shows the overlap of activity classes among different HAR datasets. ActivityNet and Charades exhibit higher overlap with HMDB51, while UCF101 and NTU RGB+D show comparatively lower similarity. Figure 1(b) presents a sample screenshot of the selected HMDB51 dataset used for the experiments.

Based on this analysis, the HMDB51 dataset is selected for our experiments. It comprises 6,766 video clips with a total size of approximately 2 GB [23]. This widely used public dataset contains 51 action categories, with each category including at least 101 video clips collected from diverse sources such as movies, YouTube, and other online platforms [23]. For this study, we focus on six specific classes relevant to medical and healthcare applications: fall_floor, walk, stand, eat, sit, and drink. These classes are chosen due to their importance in healthcare scenarios, where accurate activity recognition can provide meaningful benefits.

(a) (b)

Figure 1. Overview of the experimental datasets and dataset selection for analysis (a) class overlaps among the experimental datasets and (b) screenshot of the dataset selected for analysis

2.1. Data preprocessing

The dataset preparation involves selecting the appropriate dataset and classes. Firstly, the video data is converted to individual frames at 24 frames per second (FPS). Each frame is processed by resizing it to

224×224 pixels with RGB color channels. Then, normalizing the image data, encodes it, and finally splits the data into training, validation, and testing sets. This approach ensures that the model can effectively learn temporal dynamics and spatial features necessary for accurate action recognition across a diverse range of action categories. Subsection 3.3 meticulously details the data processing steps, which are visually represented.

The processed video clips in the HMDB51 dataset do not require any low-level filtering or enhancement, indicating that the dataset is ready for direct training of the LSTM network [24]. The videos are originally in .avi format and are converted into .jpg images for processing. The distribution shows a sharp drop after 60 frames, similar to other classes. The original dataset contains 1,258 videos across six classes (drink, eat, fall_floor, sit, stand, and walk). These videos have been converted into 18,785 images, extracted frame by frame. The images are then divided into training (75%: 14,088 images), validation (15%: 2,819 images), and test sets (10%: 1,878 images).

3. PROPOSED EfficientNetB0ConvLSTM MODEL

The proposed hybrid method, named EfficientNetB0ConvLSTM, integrates the strengths of EfficientNetB0 and LSTM networks. EfficientNetB0, a CNN architecture, is employed for feature extraction. The LSTM network is then used to capture temporal dependencies in the data, enhancing the model's performance.

3.1. EfficientNetB0 system architecture

This study proposed the EfficientNetB0 model because it is the smallest and least computationally demanding model in the EfficientNet family. Higher versions, such as EfficientNetB1, B2, B3, and B7, contain more parameters and demand more processing power and memory [25]. Additionally, the EfficientNetB0 incorporates the Swish activation function that is defined by (1).

$$f_{Swish} = \frac{1}{1+e^{-\beta x}} \tag{1}$$

Here, x is the input to the activation function, β is a hyperparameter that controls the smoothness of the swish activation function (σ) during the training of the CNN. The output layer is defined as (2) representing the computation performed by a neural network layer during the forward pass. This process allows the neural network to learn and represent complex relationships in the data, enabling it to make accurate predictions or extract meaningful features [26].

$$Y = \sigma(F * W + b) \tag{2}$$

Where Y is the output feature map, F is the input feature map, and W is the weighted matrix. The bias vector b is an additional parameter that allows the model to fit the data better by providing each neuron with a trainable constant that is added to the output of the weighted summation of inputs [27]. And the σ denotes the Swish activation function.

3.2. Convolutional long short-term memory model

The ConvLSTM architecture is a variant of the traditional LSTM that incorporates convolutional operations for processing spatial information in sequential data [28]. It is presented in Figure 2 by a control flow diagram, where input (i_t), forgets (f_t), output gates (o_t), and the (C_t) are demonstrated. The (C_t) is updated using the input and forget gates, along with the candidate (C_t) [29], [30]. For each gate, values (3) to (6) are given as follows:

$$i_t = \sigma(w_{ii} * X_t + w_{hi} * h_{t-1} + b_{ii} * b_{hi}) \tag{3}$$

$$f_t = \sigma(w_{if} * X_t + w_{hf} * h_{t-1} + b_{if} * b_{hf}) \tag{4}$$

$$o_t = \sigma(w_{io} * X_t + w_{ho} * h_{t-1} + b_{io} * b_{ho}) \tag{5}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(w_{ic} * X_t + w_{hc} * h_{t-1} + b_{ic} * b_{hc}) \tag{6}$$

Here, w_{ii}, w_{if}, w_{io} and w_{ic} are weights for the input-to-gate connections. w_{hi}, w_{hf}, w_{ho} and w_{hc} are weights for the hidden-to-gate connections. b_{ii}, b_{if}, b_{io}, and b_{ic} are the biases vector for the input-to-gate

connections. b_{hi}, b_{hf}, b_{ho}, and b_{hc} are the biases vector for the hidden-to-gate connections. Also, X_t represents the input at time step *t*, h_{t-1} and C_{t-1} denotes the hidden state and the cell state from the previous time step, respectively [30]. These mechanisms enable ConvLSTM to capture spatial and temporal dependencies in sequence data. The hidden state is computed using the output gate and the updated cell state, which is a crucial part of the LSTM and ConvLSTM architectures, representing the calculation of the hidden state h_t at time step *t*, as expressed in (7).

$$h_t = o_t \odot tanh\ (\ C_t) \tag{7}$$

This hidden state is then propagated to the subsequent time step. These equations represent how information flows through the ConvLSTM cell at each step. By incorporating convolutional operations, the model effectively captures spatial dependencies within sequential data, making ConvLSTM well-suited for tasks involving spatiotemporal patterns, such as video processing and HAR [31].

3.3. The proposed EfficientNetB0ConvLSTM model

The combination of the EfficientNetB0 architecture and the ConvLSTM network, here referred to as EfficientNetB0ConvLSTM. The proposed EfficientNetB0ConvLSTM is a sophisticated deep-learning architecture that integrates EfficientNetB0 as a feature extractor with a ConvLSTM layer for sequential data processing. EfficientNetB0, known for its efficiency and scalability, is employed to extract rich spatial features from input images. The extracted feature map is then passed to the ConvLSTM layer, which captures both spatial and temporal dependencies, making it suitable for tasks involving sequential data [32]. The overall processing pipeline of the proposed EfficientNetB0ConvLSTM model is mathematically formulated in (8) to (10).

– Feature extraction with EfficientNetB0: EfficientNetB0, a CNN model, extracts feature from input image I and is fed into EfficientNetB0, producing a feature map F.

$$F = EfficientNetB0\ (I) \tag{8}$$

– Feature map to ConvLSTM: the feature map F is reshaped to match the input requirements of the ConvLSTM layer. Here, ConvLSTM, a convolutional extension of LSTM, processes the sequential data and captures temporal dependencies along with spatial information [29]. Where F_{t-1} is the feature map, specifically, h_{t-1} and C_{t-1} represent the hidden state and cell state carried over from the previous time step, respectively, and h_t is the updated state at the current step t.

$$\{h_t, C_t\} = ConvLSTM\ (F_{t-1},\ h_{t-1}, C_{t-1}) \tag{9}$$

– Classification: the output state h_t from the final ConvLSTM layer is passed to a fully connected (dense) layer for classification. This involves multiplying by weights (w) and then adding biases (b).

$$\hat{y} = softmax\ \{(w * h_t) + b)\} \tag{10}$$

Here, w are the weights and b are the biases of the dense layer, and $\hat{y}$ is the predicted probability distribution over the classes. The EfficientNetB0ConvLSTM model block integrates EfficientNetB0 for efficient feature extraction with ConvLSTM for sequence modeling, offering a powerful combination for tasks that require processing both spatial and temporal information effectively [33].

Figure 2 presents the structure of the EfficientNetB0ConvLSTM model. The output features from the EfficientNetB0 are then fed into ConvLSTM layers to capture temporal dependencies across frames in a video or sequences in data. Each cell in a ConvLSTM layer performs convolution instead of the matrix multiplication used in standard LSTM cells, which makes it suitable for tasks involving spatial information. Moreover, the combination of EfficientNetB0 and ConvLSTM provides a powerful and efficient solution for video classification, leveraging the spatial processing power of CNNs and the temporal modeling capability of RNNs [34].

The proposed methodology is demonstrated in Figure 3, which is the comprehensive approach to developing a deep learning model for healthcare activity classification, emphasizing the importance of dataset preparation, feature extraction using EfficientNetB0, and thorough evaluation. The process begins with the collection of video data showing various activities such as drinking, eating, walking, sitting, and falling down. Then extract frames from the videos. This involves converting video sequences into a series of images. We preprocess the data by normalizing, resizing, and augmenting the images. Then split the data into training, validation, and test sets to ensure the model is trained and evaluated properly. In our study, we use

the EfficientNetB0 model as a feature extractor to extract spatial features from the images. The spatial features are fed into ConvLSTM to capture temporal dependencies. Combined features from the EfficientNetB0ConvLSTM block are used to classify the activity. The model's performance is assessed using evaluation metrics, including accuracy, precision, and recall, along with the receiver operating characteristic (ROC) curve to measure its overall effectiveness. The trained model is deployed for real-time testing, where it processes live video feeds to classify activities in real-time. This process provides a systematic approach to using a deep learning model for monitoring and classifying healthcare activities.

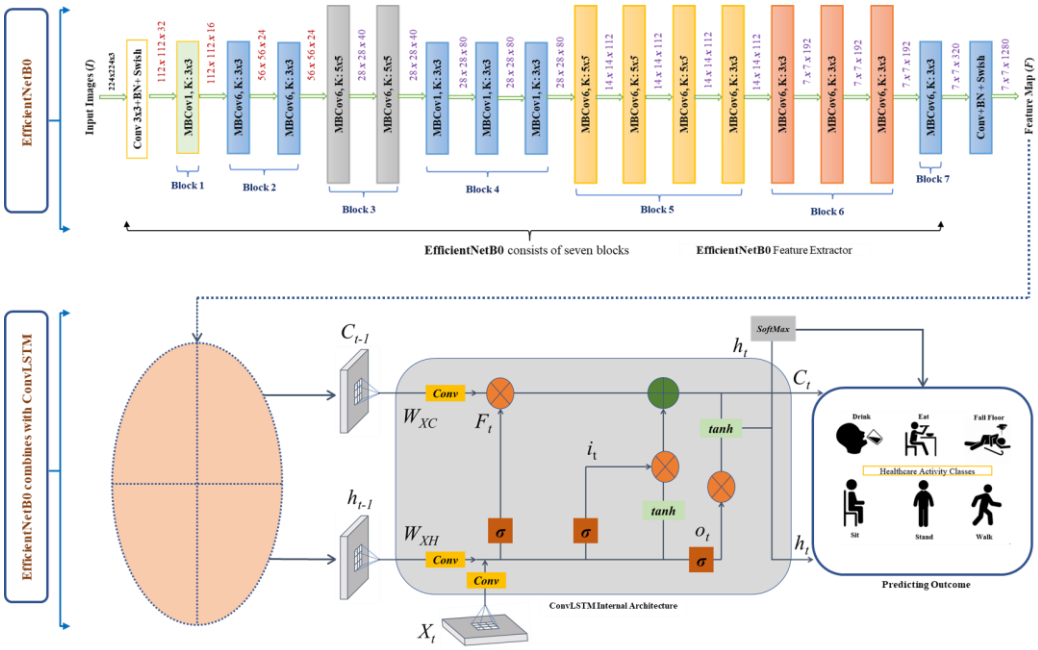


Figure 2. Internal architecture of the EfficientNetB0ConvLSTM model, which integrates EfficientNetB0 for spatial feature extraction and ConvLSTM for modeling temporal dependencies in video sequences

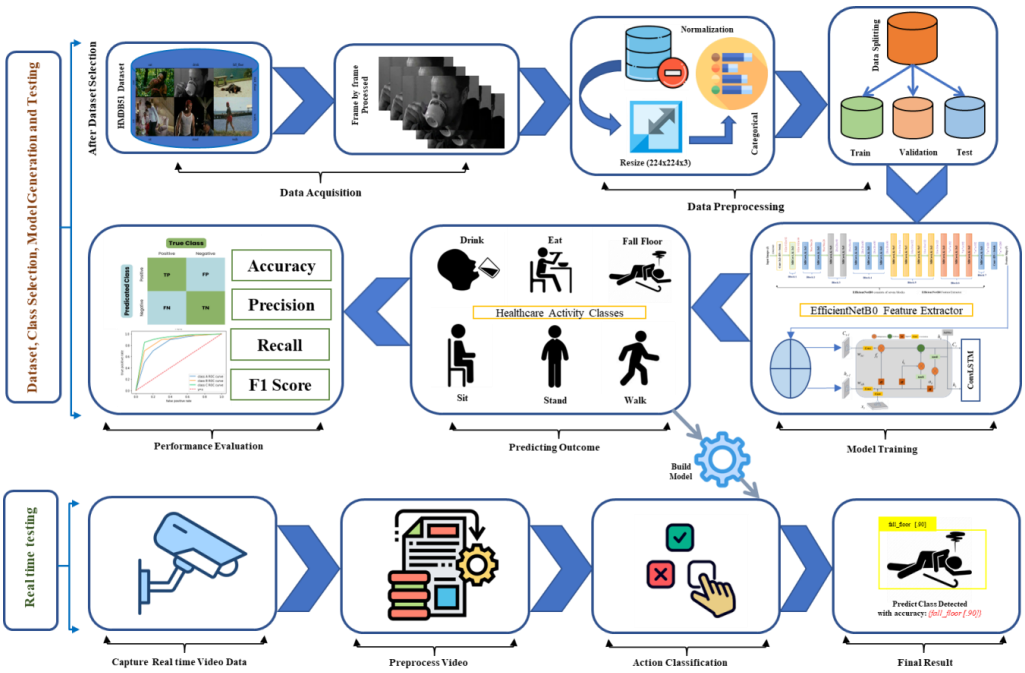


Figure 3. The overview of the proposed framework

In the proposed model, initially, it preprocesses the videos by converting them into sequences of frames and organizing them into a data frame. Next, we split the dataset into training, validation, and testing sets in a 75:15:10 ratio. We employ the EfficientNetB0 model for feature extraction, feeding these features into a ConvLSTM model, thereby integrating both models into the EfficientNetB0ConvLSTM architecture. This combines the model that refines the features and performs classification. During training, we use a tuning technique over multiple epochs, aggregating batch-oriented predictions for both training and testing phases to calculate the accuracy score. We evaluate the model's performance using model evaluation metrics such as the confusion matrix, accuracy, and loss curves, and present the final results in terms of accuracy.

The system is evaluated in real time, allowing users to input any video for activity recognition. The model processes the video frames by resizing and normalizing them before predicting activity classes using an EfficientNetB0ConvLSTM model. For each frame, the system draws a bounding box (yellow color) around the detected region of interest where the activity is occurring. It then annotates the bounding box with the predicted activity name and with its accuracy level.

4. EXPERIMENTAL RESULT

This section presents the experimental findings once the hardware and software have been configured. The experiments aim to tackle the challenge of accurate activity detection. This is achieved by training and testing deep learning architectures and applying the proposed model to the HAR challenge.

4.1. Experimental setup

Efficient training of the model largely depends on the underlying hardware configuration. In this study, we utilize an AMD Ryzen 9 5900X 12-core processor with a base clock speed of 4.20 GHz, running on a 64-bit system with Python 3.9.13 and CUDA 11.0. The setup includes an NVIDIA® GeForce RTX 3060 GAMING PC GPU with 6 GB of dedicated memory, 64 GB of RAM, and a 4 TB SSD to support large-scale data storage and processing. This configuration supports the computational demands of deep learning models and high-level interface libraries such as TensorFlow [35]. The proposed model is built using Python, Keras, TensorFlow, and additional libraries, and demonstrates robust performance in HAR.

4.2. Result analysis

Figure 4 shows the performance of the proposed model is evaluated using two primary metrics: the loss function and accuracy. These metrics are monitored during both the training and validation phases. Figure 4(a) illustrates the model’s behavior during training, where both training and validation accuracy improve progressively with an increasing number of epochs. Similarly, the corresponding loss curves show a consistent downward trend. Figure 4(b) further confirms effective learning, as the decreasing loss indicates that the model is gradually reducing the discrepancy between predicted outputs and true labels. The overall data summary is provided in Table 3.

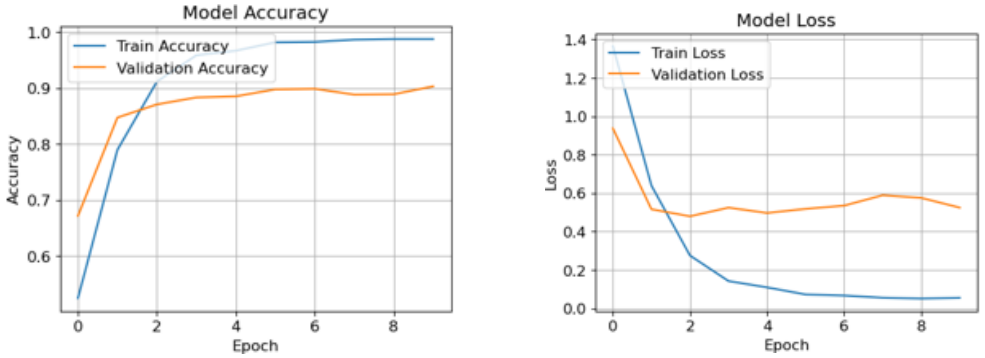
The training and validation metrics reveal a training accuracy of 89.22% and a validation accuracy of 75.63%, with a training loss of 0.5842±0.101 and a validation loss of 3.1545±0.0486. The model's performance in human activity classification is further illustrated by the normalized confusion matrix, presented in Figure 5. This figure illustrates the classification performance of the proposed EfficientNetB0ConvLSTM model using confusion matrix analysis. Figure 5(a) presents the confusion matrix, where diagonal elements indicate correctly classified samples and off-diagonal elements represent misclassifications. Higher accuracy is observed for the “drink”, “eat”, and “walk” classes. Figure 5(b) shows the normalized confusion matrix in percentage form, providing clearer class-wise performance analysis. The “walk” and “drink” classes achieve the highest recognition rates, while “fall_floor” shows comparatively lower accuracy. This matrix provides a granular view of class-specific accuracy, where the diagonal elements represent correctly classified instances, and off-diagonal elements indicate misclassifications. Notably, the classes “drink”, “eat”, and “walk” exhibit exceptional accuracy, while the remaining classes demonstrate comparatively lower performance. Table 3 presents a summary of the standard classification metrics for the HMDB51 dataset, with an overall average accuracy of 89.22% achieved in our experiments.

Figure 6 presents the evaluation analysis of the proposed model using optimizer performance and precision-recall (PR) analysis. Figure 6(a) compares validation loss across Adam, stochastic gradient descent (SGD), and root mean square propagation (RMSProp) optimizers. Adam achieves faster convergence and lower validation loss, demonstrating superior performance. Figure 6(b) illustrates the PR curves for different activity classes. The “drink” and “walk” classes achieve better PR performance, while “fall_floor” shows relatively lower performance. The results indicate that the Adam optimizer performs better than both SGD and RMSProp. Overall, Adam emerges as the most effective optimizer among the three: Adam, SGD, and RMSProp considered in this study. This superior performance can be attributed to its ability to combine the

benefits of RMSProp and momentum, which enables faster and more stable convergence [36]. In addition, Adam adaptively adjusts the learning rate during training while leveraging historical gradient information to accelerate the learning process, making it highly efficient for deep learning applications [37]. Furthermore, it achieves the highest validation accuracy and the lowest validation loss among all tested optimizers. Therefore, the Adam optimizer is selected for this study. The PR curve is a crucial tool for understanding and optimizing the performance of classification models, especially in cases where the positive class is much less frequent than the negative class. It provides detailed insights into the PR trade-off, aids in threshold selection, and facilitates comparative analysis of models. From Figure 6(b), we can understand that the "drink" and "walk" classes have the best performance, staying closer to the top right corner, indicating high precision and recall. The "fall_floor" class shows a noticeable drop in performance as recall increases, indicating a trade-off with precision. Other classes like "eat", "sit", and "stand" have varying performances, with precision decreasing as recall increases.

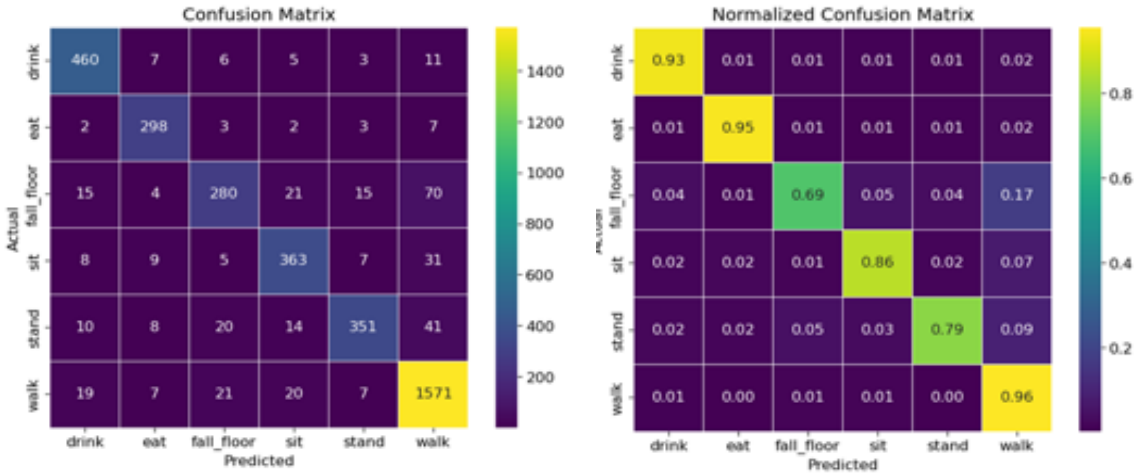
We apply different activation functions to our experimental model and plot their performances in a single chart, as presented in Figure 7. From this analysis, we observe that the rectified linear unit (ReLU) activation function performs the best. For this purpose, we employ the ReLU activation function in our model. This function is mathematically defined in (11). Figure 7 compares different activation functions in terms of accuracy and loss performance. Figure 7(a) shows that the ReLU activation function achieves the highest classification accuracy. Figure 7(b) demonstrates that ReLU also provides lower loss and more stable convergence compared to other activation functions. Therefore, ReLU is selected for the proposed model.

$$ReLU\ (r) = max(0, r); \begin{cases} r, & r > 0 \\ 0, & r \leq 0 \end{cases} \tag{11}$$



(a) (b)

Figure 4. Performance evaluation of (a) model accuracy and (b) model loss

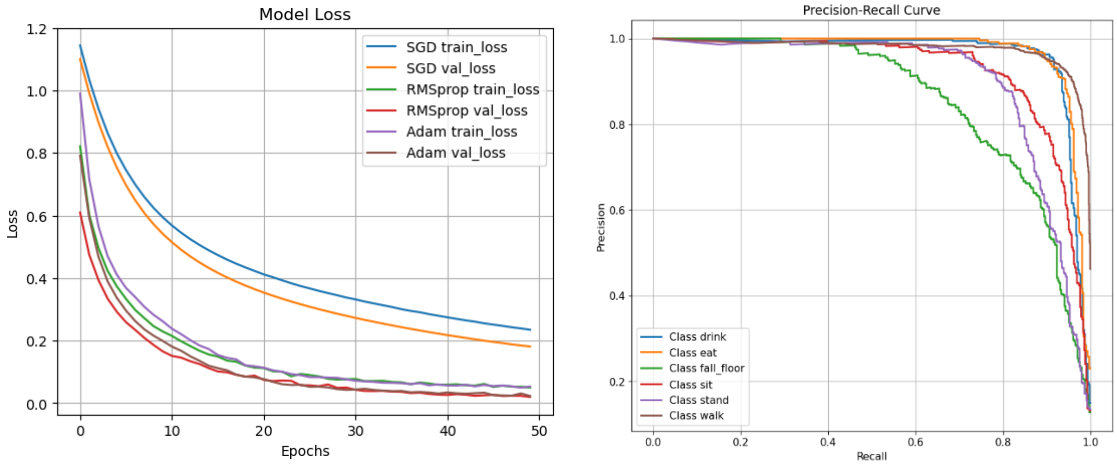


(a) (b)

Figure 5. Performance visualization of (a) confusion matrix and (b) its normalized representation

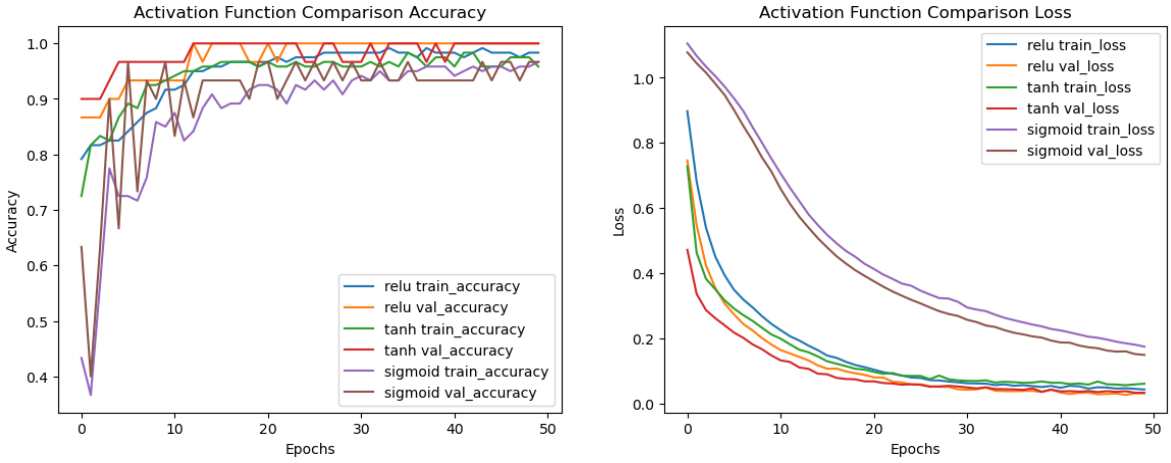
Table 3. Classification result on the HMDB51 dataset

Class levels	Precision	Recall	F1-score	Support
drink	0.89	0.93	0.91	492
eat	0.89	0.95	0.92	315
fall_floor	0.84	0.69	0.76	405
sit	0.85	0.86	0.86	423
stand	0.91	0.79	0.85	444
walk	0.91	0.96	0.93	1645
Accuracy			**0.89**	3724
Macro avg	0.88	0.86	0.87	3724
Weighted avg	0.89	0.89	0.89	3724



(a) (b)

Figure 6. Evaluation metrics (a) PR curve and (b) model loss across optimizer functions



(a) (b)

Figure 7. Comparison of different activation functions of (a) accuracy comparison and (b) loss comparison

Table 4 illustrates the performance metrics for various models tested on the HMDB51 dataset for human action recognition. The results highlight that our proposed model, EfficientNetB0ConvLSTM, attains the highest accuracy of 89.22%. Among other notable models, BiLSTM achieves 76.30% accuracy, and ViT+PBiLSTM+DSMHA achieves 78.62%. The STM, attention-based LSTM and 3D CNN, and EfficientNet models perform with an accuracy of 80.40%, 87.98%, and 88.70% respectively, in contrast, earlier and other baseline models such as VicTR (ViT-B/16), SVT (Fine-tune), SVT (Linear), and LSTM demonstrate relatively lower performance, achieving accuracies of 51.00%, 67.28%, 57.8%, and 73.75%, separately. However, with the EfficientNetB0ConvLSTM model, we propose only the selected 6 classes,

which reach the highest performance, with an accuracy of 89.22%, precision of 88.12%, recall of 86.54%, and an F1-score of 87.96%. We also tested other well-known models; the LSTM model achieves an accuracy of 76.44%, with a precision of 75.96%, a recall of 76.42%, and an F1-score of 78.18%. The ConvLSTM model performs an accuracy of 70.35%, with a precision of 71.07%, a recall of 73.33%, and an F1-score of 69.28%. The EfficientNetB0 model performs with an accuracy of 72.57%, precision of 74.24%, recall of 72.57%, and an F1-score of 75.45%. Meanwhile, the EfficientNet model attains an accuracy of 76.75%, with a precision of 74.48%, a recall of 71.73%, and an F1-score of 73.65%. Table 5 presents the detailed results, clearly demonstrating that the proposed model achieves superior performance across all evaluation metrics.

Table 4. A quantitative comparison of the proposed method with state-of-the-art action recognition techniques on the HMDB51 dataset, organized by publication year

Citations	Year	Accuracy (%)
Xiao *et al.* [7]	2022	75.91
Ranasinghe *et al.* [8]	2022	57.82
		67.21
Chen *et al.* [9]	2022	51.03
Xing *et al.* [10]	2023	59.71
		68.22
Ullah and Munir [11]	2023	79.38
Dastbaravardeh *et al.* [12]	2024	77.29
Barr *et al.* [13]	2024	88.70
Hussain *et al.* [14]	2024	76.30
Hassan *et al.* [15]	2024	81.20
Sun *et al.* [38]	2024	80.40
Proposed method	2025	89.22

Table 5. Comparative performance of tested models on the HMDB51 dataset in our experiments

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LSTM	76.44	75.96	76.42	78.18
ConvLSTM	70.35	71.07	73.33	69.28
EfficientNetB0	72.57	74.24	72.57	75.45
EfficientNet (B0-B7)	76.75	74.48	71.73	73.65
EfficientNetB0ConvLSTM [Proposed]	89.22	88.12	86.54	87.96

5. REAL-TIME IMPLEMENTATION AND TESTING

The performance of the proposed HAR system in the laboratory experiment is satisfactory. However, there are always some differences between laboratory and real-time scenarios. The system utilizes an Imou Ranger 2 WiFi closed-circuit camera, designed for indoor use, providing high-resolution video monitoring and ensuring comprehensive surveillance and security for the home environment. This configures the camera to operate at 24 FPS. To minimize delays from wireless networks, we use a USB cable for a direct connection. The camera is linked to a computer where our program is installed. This setup is crucial for achieving accurate and timely observations and analyses in the study. The experimental results presented in this paper are measured at a 360-degree angle, and the camera is positioned at the Digipod TR-462 camera tripod stand for capturing video data, and passed to our proposed model to detect human activity with its predicted accuracy.

5.1. Insight and evaluation

The performance of the proposed algorithm in real-time is illustrated in Figure 8. The performance is observed over seconds, where a patient/human does within these classes, the proposed algorithm effectively identifies the target incidents consequently. Expect these six classes; this model can't predict the class level. Based on the ratio of the laboratory experiment and real-time activity identification, a performance score is generated. The performance score is calculated using the equation defined in (12). Where, n =1, 2, 3, 4, 5, 6 (1= fall_floor, 2= walk, 3= stand, 4= eat, 5= sit, 6= drink).

$$Score_n = \frac{Real\_time\_Experiment_n}{Laboratory\_Experiment_n} \quad (12)$$

The performance scores for the experimental incidents, based on real-time observations, are listed in Table 6. The proposed algorithm, EfficientNetB0ConvLSTM, is a system that is effective in detecting human activity in the medical sector based on its performance score. The performance is a little bit degraded when

real-time testing is captured. For real-time testing, we collect data from “Majumder Clinic”, located in Dayarampur, Natore, Bangladesh. The clinic's authority granted permission for the collection of real-time data from its patients.

Figure 8. Performance on the real-time testing with testing accuracy score left: “stand”, middle: “eat”, “drink”, “fall_floor”, “sit” and right: “walk”

Table 6. Class-wise performance evaluation based on observational analysis

Actual class	Real-world observation	Laboratory accuracy (%)	Real-time accuracy (%)	Calculate score (%)	Real-time Loss (%)
fall_floor	Person falling to the ground	89.00	87.00	97.75	2.00
walk	Person walking	95.00	91.00	95.79	4.00
stand	Person standing still	79.00	77.00	97.47	2.00
eat	Person eating	94.00	90.00	95.74	4.00
sit	Person sitting	86.00	82.00	95.35	4.00
drink	Person drinking	92.00	88.00	95.65	4.00
	Average score	89.22	85.83	96.29	3.33

6. LIMITATION AND FUTURE RESEARCH SCOPE

The proposed system has certain limitations, which highlight potential directions for future research. It struggles with subjects that are either too close (15 cm) or too far (25 m) from the camera, affecting performance. The system has only been tested in well-lit conditions, leaving night vision performance unexamined. It currently handles only six medical classes due to data constraints, indicating the need for more diverse datasets and class expansion. Environmental factors, such as lighting and clutter, can impact accuracy, and the combined use of deep learning models increases computational demands, requiring robust hardware for real-time operation. Addressing these limitations could enhance system robustness and versatility.

7. CONCLUSION

In this study, a hybrid EfficientNetB0-ConvLSTM framework was proposed for HAR, aiming to reduce computational complexity while enhancing overall performance. The model integrates EfficientNetB0 for spatial feature extraction and ConvLSTM for temporal classification, achieving an accuracy of 89.22%, precision of 88.12%, recall of 86.54%, and an F1-score of 87.96% across six medically relevant classes from the HMDB51 dataset. The approach eliminates the need for additional equipment, relying on standard CCTV or webcams for activity recognition, making it suitable for healthcare and security applications. Real-time experiments demonstrated the system's adaptability in dynamic environments, confirming its reliability in monitoring activities with high confidence scores. Despite its strong performance, limitations such as sensitivity to subject position, environmental conditions, and night-vision challenges were identified, highlighting opportunities for further optimization. These insights pave the way for future advancements in affordable, efficient, and accessible HAR systems.

ACKNOWLEDGMENTS

The authors would like to thank the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh, for supporting this research through the ICT fellowship in M.Phil. program. The authors also acknowledge the support of the Department of Computer Science and Engineering, University of Rajshahi, Bangladesh, and Bangladesh Army University of Engineering and Technology (BAUET). In addition, the authors express their gratitude to Majumder Clinic, Dayarampur, Natore, Bangladesh, for permitting real-time data collection for experimental analysis.

FUNDING INFORMATION

This research was funded by the ICT Division, Ministry of Posts, Telecommunications, and Information Technology, Bangladesh, under the ICT Fellowship in my M.Phil. Program (Grant No: 56.00.0000.052.33.005.21-7, Tracking No: 22FS15306) with support from the University of Rajshahi.

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
Subrata Kumer Paul	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Rakhi Rani Paul		✓				✓		✓	✓	✓	✓	✓		
Md. Ekramul Hamid	✓		✓	✓			✓			✓	✓		✓	✓
Md. Rafiqul Islam (Rafiq)		✓				✓			✓	✓	✓			

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

The authors declare no conflict of interest.

INFORMED CONSENT

Informed consent was obtained from the participants involved in the real-time data collection process.

ETHICAL APPROVAL

The study was conducted following relevant institutional and national ethical guidelines with permission from the concerned clinical authority.

DATA AVAILABILITY

The HMDB51 dataset is publicly available in the HMDB51 dataset repository at https://serre.lab.brown.edu/hmdb51.html. Additional experimental data are available from the corresponding author, [SKP], upon reasonable request.

REFERENCES

[1] S. Hurtado, J. G.-Nieto, A. Popov, and I. N.-Delgado, "Human activity recognition from sensorised patient´s data in healthcare: a streaming deep learning-based approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 23–37, 2023, doi: 10.9781/ijimai.2022.05.004.

[2] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, and Y. Wang, "Abnormal event detection using deep contrastive learning for intelligent video surveillance system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5171–5179, Aug. 2022, doi: 10.1109/TII.2021.3122801.

[3] R. R. Paul, S. K. Paul, and M. E. Hamid, "A 2D convolution neural network based method for human emotion classification from speech signal," in *Proceedings of 2022 25th International Conference on Computer and Information Technology, ICCIT 2022*, 2022, pp. 72–77, doi: 10.1109/ICCIT57492.2022.10054811.

[4] F. S. Abuhoureyah, Y. C. Wong, and A. S. M. Isira, "WiFi-based human activity recognition through wall using deep learning," *Engineering Applications of Artificial Intelligence*, vol. 127, Jan. 2024, doi: 10.1016/j.engappai.2023.107171.
[5] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human action recognition systems: a review of the trends and state-of-the-art," *IEEE Access*, vol. 12, pp. 36372–36390, 2024, doi: 10.1109/ACCESS.2024.3373199.
[6] S. K. Paul *et al.*, "An Adam-based CNN and LSTM approach for sign language recognition in real time for deaf people," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 499–509, 2024, doi: 10.11591/eei.v13i1.6059.
[7] J. Xiao *et al.*, "Learning from temporal gradient for semi-supervised action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3252–3262, doi: 10.1109/CVPR52688.2022.00325.
[8] K. Ranasinghe, M. Naseer, S. Khan, F. S. Khan, and M. S. Ryoo, "Self-supervised video transformer," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 2864–2874, 2022, doi: 10.1109/CVPR52688.2022.00289.
[9] S. Chen *et al.*, "AdaptFormer: adapting vision transformers for scalable visual recognition," *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.
[10] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y. G. Jiang, "SVFormer: semi-supervised video transformer for action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023, vol. 2023-June, pp. 18816–18826, doi: 10.1109/CVPR52729.2023.01804.
[11] H. Ullah and A. Munir, "Human activity recognition using cascaded dual attention CNN and bi-directional GRU framework," *Journal of Imaging*, vol. 9, no. 7, 2023, doi: 10.3390/jimaging9070130.
[12] E. Dastbaravardeh, S. Askarpour, M. S. Anari, and K. Rezaee, "Channel attention-based approach with autoencoder network for human action recognition in low-resolution frames," *International Journal of Intelligent Systems*, vol. 2024, pp. 1–22, Jan. 2024, doi: 10.1155/2024/1052344.
[13] J. B.-Barr, B. Fernando, and D. Rajan, "Activation control of vision models for sustainable AI systems," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 7, pp. 3470–3481, Jul. 2024, doi: 10.1109/TAI.2024.3372935.
[14] A. Hussain *et al.*, "Shots segmentation-based optimized dual-stream framework for robust human activity recognition in surveillance video," *Alexandria Engineering Journal*, vol. 91, pp. 632–647, Mar. 2024, doi: 10.1016/j.aej.2023.11.017.
[15] N. Hassan, A. S. M. Miah, and J. Shin, "A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition," *Applied Sciences*, vol. 14, no. 2, Jan. 2024, doi: 10.3390/app14020603.
[16] W. Wang *et al.*, "A spatiotemporal and motion information extraction network for action recognition," *Wireless Networks*, vol. 30, no. 6, pp. 5389–5405, Aug. 2024, doi: 10.1007/s11276-023-03267-y.
[17] H. Quan, Y. Hu, and A. Bonarini, "POLIMI-ITW-S: a large-scale dataset for human activity recognition in the wild," *Data in Brief*, vol. 43, Aug. 2022, doi: 10.1016/j.dib.2022.108420.
[18] Y. Zeng, Y. Zhong, C. Feng, and L. Ma, "UniMD: towards unifying moment retrieval and temporal action detection," in *Computer Vision – ECCV 2024: 18th European Conference, Proceedings, Part XLVI*, Milan, Italy, pp. 286–304, doi: 10.1007/978-3-031-72952-2_17.
[19] M. Zakariah and A. Alnuaim, "Recognizing human activities with the use of convolutional block attention module," *Egyptian Informatics Journal*, vol. 27, Sep. 2024, doi: 10.1016/j.eij.2024.100536.
[20] A. Z. S. Yii, K. H. Lim, and C. W. R. Chiong, "Review of three-dimensional human action recognition," in *2024 International Conference on Green Energy, Computing and Sustainable Technology, GECOST 2024*, Jan. 2024, pp. 349–353, doi: 10.1109/GECOST60902.2024.10474856.
[21] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "MetaVD: a meta video dataset for enhancing human action recognition datasets," *Computer Vision and Image Understanding*, vol. 212, Nov. 2021, doi: 10.1016/j.cviu.2021.103276.
[22] Center for Research in Computer Vision, "UCF101-action recognition data set, University of Central Florida." crcv.ucf.edu. Accessed: Aug. 2, 2024. [Online]. Available: https://www.crcv.ucf.edu/data/UCF101.php
[23] Serre Lab, "HMDB: a large human motion database." serre.lab.brown.edu. Accessed: Aug. 2, 2024. [Online]. Available: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/
[24] L. P. O. Paula, N. Faruqui, I. Mahmud, M. Whaiduzzaman, E. C. Hawkinson, and S. Trivedi, "A novel front door security (FDS) algorithm using GoogleNet-BiLSTM hybridization," *IEEE Access*, vol. 11, pp. 19122–19134, 2023, doi: 10.1109/ACCESS.2023.3248509.
[25] R. Raza *et al.*, "Lung-EffNet: lung cancer classification using EfficientNet from CT-scan images," *Engineering Applications of Artificial Intelligence*, vol. 126, Nov. 2023, doi: 10.1016/j.engappai.2023.106902.
[26] A. Masood and K. Ahmad, "A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: fundamentals, application and performance," *Journal of Cleaner Production*, vol. 322, Nov. 2021, doi: 10.1016/j.jclepro.2021.129072.
[27] R. Daghigh, S. A. Arshad, K. Ensafjoee, and N. Hajialigol, "A data-driven model for a liquid desiccant regenerator equipped with an evacuated tube solar collector: random forest regression, support vector regression and artificial neural network," *Energy*, vol. 295, May 2024, doi: 10.1016/j.energy.2024.130932.
[28] A. V. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: advancements, challenges, and future directions," *Information Fusion*, vol. 105, May 2024, doi: 10.1016/j.inffus.2023.102218.
[29] S. K. Paul, A. S. M. Miah, R. R. Paul, M. E. Hamid, J. Shin, and M. A. Rahim, "IoT-based real-time medical-related human activity recognition using skeletons and multi-stage deep learning for healthcare," *Computers, Materials and Continua*, vol. 84, no. 2, pp. 2513–2530, 2025, doi: 10.32604/cmc.2025.063563.
[30] S. K. Paul and R. R. Paul, "Speech command recognition system using deep recurrent neural networks," in *2021 5th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2021*, 2021, pp. 1–6, doi: 10.1109/ICEEICT53905.2021.9667795.
[31] N. Dündar, A. S. Keçeli, A. Kaya, and H. Sever, "A shallow 3D convolutional neural network for violence detection in videos," *Egyptian Informatics Journal*, vol. 26, Jun. 2024, doi: 10.1016/j.eij.2024.100455.
[32] M. T. R., M. Gupta, A. T. A., V. Kumar V, O. Geman, and D. Kumar V., "An XAI-enhanced efficientNetB0 framework for precision brain tumor detection in MRI imaging," *Journal of Neuroscience Methods*, vol. 410, 2024, doi: 10.1016/j.jneumeth.2024.110227.
[33] M. S. Dizaji, Z. Mao, and M. Haile, "A hybrid-attention-ConvLSTM-based deep learning architecture to extract modal frequencies from limited data using transfer learning," *Mechanical Systems and Signal Processing*, vol. 187, Mar. 2023, doi: 10.1016/j.ymssp.2022.109949.
[34] 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.
[35] H. Li, G. K. Rajbahadur, and C. P. Bezemer, "Studying the impact of TensorFlow and PyTorch bindings on machine learning software quality," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 1, Jul. 2024, doi: 10.1145/3678168.

[36] I. U. W. Mulyono, Y. Kusumawati, A. Susanto, C. A. Sari, H. M. M. Islam, and M. Doheir, "Hiragana character classification using convolutional neural networks methods based on Adam, SGD, and RMSProps optimizer," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 467–476, 2024, doi: 10.15294/sji.v11i2.2313.
[37] H. Sun *et al.*, "AdaSAM: boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks," *Neural Networks*, vol. 169, pp. 506–519, Jan. 2024, doi: 10.1016/j.neunet.2023.10.044.
[38] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: spatiotemporal and motion encoding for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2000–2009, doi: 10.1109/ICCV.2019.00209.

BIOGRAPHIES OF AUTHORS

Subrata Kumer Paul completed his B.Sc. Engineering, M.Sc. Engineering, and Master of Philosophy (M.Phil.) in Computer Science and Engineering from the University of Rajshahi in 2016, 2019, and 2025, respectively. He is currently serving as an assistant professor in the Department of Computer Science and Engineering at the Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore-6431, Bangladesh. He is presently pursuing his Ph.D. at the same institution. His research interests include HR, deep learning, artificial intelligence, movement disorders (ASD, AD, and PD), and signal processing. He has published more than 30 research articles in reputed international journals, conferences, and book chapters. In addition, he has actively participated in numerous symposiums, poster presentations, workshops, and seminars related to research and innovation. He has been awarded a fellowship from the ICT Division under the Ministry of Posts, Telecommunications, and Information Technology of Bangladesh. He has also received best paper and best researcher awards in recognition of his outstanding research contributions. He is a graduate member of IEEE. He can be contacted at email: sksubrata96@gmail.com.

Rakhi Rani Paul graduated with her B.Sc. and M.Sc. Engineering from University of Rajshahi, in Computer Science and Engineering in 2017 and 2018, respectively. Now, she is working as an assistant professor at the Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore-6431, Bangladesh. Her research fields are deep learning, data mining, and speech signal processing. She has actively participated in numerous symposiums, poster presentations, workshops, and seminars on research-related topics. She is presently pursuing her Ph.D. at Rajshahi University. She has published more than 25 international journal/book chapter/conference papers. She can be contacted at email: rakhipaul.cse@gmail.com.

Dr. Md. Ekramul Hamid received his B.Sc. and M.Sc. degrees in Applied Physics and Electronics from the University of Rajshahi. Later on, received an M.Cs. degree from Pune University, India, and a Ph.D. degree from Shizuoka University, Japan. He is currently working as a professor at the Department of Computer Science and Engineering, University of Rajshahi, Bangladesh. He has published more than 70 international journal/conference papers. In 2026, he secured the ICSETEP project and is serving as the principal investigator (PI) of the project. He is a recipient of the Monbukagakusho scholarship, JASSO Fellowship, and NIST fellowship for his contribution to Science and Technology. He worked as a faculty member at King Khalid University, KSA, in 2010-11 and as a visiting researcher at Shizuoka University, Japan, in 2012, 2014, and 2017, respectively. He worked as the chairman of the CSE Department from September 2011 to June 2015 and as dean of the Faculty of Engineering from April 2018 to November 2021 at the University of Rajshahi. His research interests include audio signal processing, speech enhancement, machine learning, and image processing. He can be contacted at email: ekram_hamid@ru.ac.bd.

Md. Rafiqul Islam (Rafiq) received his B.Sc. Engineering degree in Computer Science and Engineering from the Dhaka University of Engineering and Technology (DUET), Gazipur, Bangladesh, in 2016. He is currently pursuing his M.Sc. Engineering degree in Information and Communication Technology at the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. He is working as a Lecturer with the Department of Computer Science and Engineering, Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore, Bangladesh. His research interests include machine learning, computer vision, image processing, cryptography, and network security. He has authored several research articles published in peer-reviewed national and international journals and conference proceedings. He is actively involved in teaching, research, and curriculum development. He can be contacted at rafique.cseduet@gmail.com