Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063563

ARTICLE

IoT-Based Real-Time Medical-Related Human Activity Recognition Using Skeletons and Multi-Stage Deep Learning for Healthcare

Subrata Kumer Paul[1,2], Abu Saleh Musa Miah[3,4], Rakhi Rani Paul[1,2], Md. Ekramul Hamid[2], Jungpil Shin[4,*] and Md Abdur Rahim[5]

[1]Department of Computer Science and Engineering (CSE), Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad Cantonment, Dayarampur, Natore, 6431, Bangladesh
[2]Department of Computer Science and Engineering (CSE), Rajshahi University, Rajshahi, 6205, Bangladesh
[3]Department of CSE, Bangladesh Army University of Science and Technology (BAUST), Nilphamari, Saidpur, 5311, Bangladesh
[4]School of Computer Science and Engineering, The University of Aizu, Aizuwakmatsu, 965-8580, Japan
[5]Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, 6600, Bangladesh

*Corresponding Author: Jungpil Shin. Email: jpshin@u-aizu.ac.jp

Received: 17 January 2025; Accepted: 30 April 2025; Published: 03 July 2025

ABSTRACT: The Internet of Things (IoT) and mobile technology have significantly transformed healthcare by enabling real-time monitoring and diagnosis of patients. Recognizing Medical-Related Human Activities (MRHA) is pivotal for healthcare systems, particularly for identifying actions critical to patient well-being. However, challenges such as high computational demands, low accuracy, and limited adaptability persist in Human Motion Recognition (HMR). While some studies have integrated HMR with IoT for real-time healthcare applications, limited research has focused on recognizing MRHA as essential for effective patient monitoring. This study proposes a novel HMR method tailored for MRHA detection, leveraging multi-stage deep learning techniques integrated with IoT. The approach employs EfficientNet to extract optimized spatial features from skeleton frame sequences using seven Mobile Inverted Bottleneck Convolutions (MBConv) blocks, followed by Convolutional Long Short Term Memory (ConvLSTM) to capture spatio-temporal patterns. A classification module with global average pooling, a fully connected layer, and a dropout layer generates the final predictions. The model is evaluated on the NTU RGB+D 120 and HMDB51 datasets, focusing on MRHA such as sneezing, falling, walking, sitting, etc. It achieves 94.85% accuracy for cross-subject evaluations and 96.45% for cross-view evaluations on NTU RGB+D 120, along with 89.22% accuracy on HMDB51. Additionally, the system integrates IoT capabilities using a Raspberry Pi and GSM module, delivering real-time alerts via Twilios SMS service to caregivers and patients. This scalable and efficient solution bridges the gap between HMR and IoT, advancing patient monitoring, improving healthcare outcomes, and reducing costs.

KEYWORDS: Real-time human motion recognition (HMR); ENConvLSTM; EfficientNet; ConvLSTM; skeleton data; NTU RGB+D 120 dataset; MRHA

1 Introduction

Human Motion Recognition (HMR) systems aim to automatically identify and monitor individual or group activities, playing a crucial role in healthcare by tracking physical and medical-related activities. Among these, medical-related human activities (MRHA), such as sneezing, coughing, falling, and staggering, are vital for ensuring patient safety and well-being. However, accurate detection of MRHA remains a significant challenge. With the global elderly population expected to reach 2.1 billion by 2050, according

to the World Health Organization (WHO) and the United Nations, the demand for effective healthcare solutions is becoming increasingly critical. Many elderly individuals live alone or in care facilities where healthcare professionals are often outnumbered, leading to insufficient monitoring and increased risks. Detecting MRHA, such as falls or signs of distress, is essential to enable real-time interventions and prevent critical incidents. Falls are a particularly significant concern, causing over 646,000 deaths and 37 million severe injuries annually, as reported by the WHO. As the elderly population grows, accounting for approximately 16% of the global population by 2050, the need for automated MRHA detection systems becomes increasingly urgent. These systems enhance patient safety, provide continuous monitoring, and reduce the strain on healthcare professionals [1–4]. Medical-Related Human Activity (MRHA) recognition systems are vital for ensuring safety and providing continuous monitoring of in-home care and assisted medical facilities. These systems enable timely interventions in cases of medical emergencies, such as falls, significantly reducing mortality risk by 80% and minimizing extended hospital stays by 26% [5]. Current approaches for MRHA detection primarily use wearable sensors and vision-based methods [3], with sensors capturing acceleration changes associated with falls [6] and vision-based systems analyzing video data [7]. However, developing robust automated MRHA recognition systems is crucial to deliver timely interventions and prevent severe injuries and fatalities. In this study, we selected the NTU RGB+D 120 and HMDB51 datasets due to their diverse range of human activities, including several medically relevant motions such as falling, walking, sitting, drinking, and eating, which are crucial for healthcare applications. NTU RGB+D 120, in particular, is one of the largest and most widely used action recognition datasets.

1.1 Current Fall Detection Systems and Their Challenges

Despite advancements, existing MRHA recognition technologies face significant limitations. Wearable sensor-based systems, though effective in detecting acceleration changes, often struggle with user comfort, false positives during non-critical activities, and low compliance among elderly individuals with cognitive impairments [8]. Vision-based systems provide a non-invasive alternative but raise privacy concerns, as video feeds can compromise individual anonymity and legal safeguards, even when encoding techniques are employed to obscure clarity. Skeleton-based data, derived from pose estimation algorithms [9] or Kinect systems [10], offers a promising solution. Kinect systems are used in human activity recognition (HAR) by capturing 3D skeletal data through depth sensors and tracking key joints to build a model of human posture and movement. This approach preserves privacy by omitting identifiable information while maintaining robustness against challenges such as background noise and lighting variations. Skeleton-based data also have lower dimensionality, ensuring efficient motion representation with reduced computational costs. Recent studies have highlighted the efficacy of skeleton-based MRHA recognition. For instance, Zahan et al. [11] achieved over 94% accuracy on URFD and UPFD datasets using Graph Convolutional Networks (GCNs) combined with Convolutional Neural Networks (CNNs). Similarly, Egawa et al. [2] applied a modified GCN model to the ImVia RU-Fall dataset, reporting a 99.00% accuracy.

1.2 Emerging Datasets and Research Gaps

Two important datasets, NTU RGB+D 120 and HMDB51, include videos of medical-related activities like sneezing, coughing, sitting, and walking, which are useful for healthcare. However, only a few researchers have used these datasets for MRHA recognition, and the reported accuracy is still low. Improving MRHA recognition using these datasets can lead to systems that are more accurate, private, and flexible. Future work should focus on creating better models that work well for different populations and activity types.

1.3 Research Motivation

The growing elderly population and increasing fall rates highlight the need for accurate, efficient, and privacy-preserving medical-related human activity (MRHA) recognition systems. Falls are a major cause of injury and mortality in older adults, driving up healthcare costs and affecting quality of life. Traditional solutions like wearable sensors and vision-based systems face issues such as low accuracy, portability challenges, and privacy concerns. This study proposes a robust solution leveraging IoT and mobile technology for real-time patient monitoring. By using skeleton data to capture joint movements while ensuring privacy, the research aims to develop an advanced human motion recognition (HMR) framework for healthcare. This approach enhances safety, improves outcomes, and reduces costs, advancing modern healthcare systems.

1.4 The Goal and Scope of the Study

This study develops a real-time Medical-Related Human Activity (MRHA) recognition system using a multi-stage deep learning model and IoT integration. It features direct mobile notifications without third-party apps for fast, secure health alerts. The system is rigorously validated for accuracy, offering timely, data-driven support for improved patient care. Key contributions includes:

- **Novel Hybrid Deep Learning Model for MRHA Recognition:** We propose ENConvLSTM, a multi-stage deep learning model combining EfficientNet for spatial features and ConvLSTM for spatio-temporal integration. It addresses key HMR challenges like high computational cost, low accuracy, and poor adaptability. Using seven MBConv blocks, it enhances spatial representation and motion analysis.
- **Exceptional Performance on Benchmark Datasets:** The model is evaluated on the NTU RGB+D 120 that is presented in Fig. 1 and HMDB51 datasets, focusing on MRHA such as sneezing, falling, walking, and sitting. It achieves 94.85% accuracy for cross-subject evaluations and 96.45% for cross-view evaluations on NTU RGB+D 120, along with 89.22% accuracy on HMDB51. These results demonstrate the model's capability to handle both spatial and temporal data aspects effectively.
- **Real-Time IoT-Integrated MRHA Recognition System:** A real-time IoT system using Raspberry Pi and a GSM module with Twilio API delivers instant SMS alerts, eliminating the need for third-party apps. It recognizes 12 MRHAs (e.g., sneezing, falling, walking), enabling early diagnosis, timely intervention, and improved healthcare outcomes.

Figure 1: Sample example of "NTU RGB+D 120" dataset

The paper is organized as follows: Section 2 presents the literature review and major contributions. Section 3 describes the dataset, selection criteria, and preprocessing. The proposed methodology is

discussed in Section 4, followed by experimental results and performance evaluation in Section 5. Real-time implementation and analysis are presented in Section 6. Section 7 concludes the paper and suggests future directions.

2 Related Works

The convergence of IoT and mobile technologies has revolutionized healthcare, enabling real-time patient monitoring and diagnosis. One of the crucial areas in this domain is human motion recognition (HMR), which various methodologies have employed over the last few years. In this survey, we consider only the "NTU RGB+D 120" dataset. Cross-Subject (CS) and Cross-View (CV) are two accuracy evaluation methods. Plizzari et al. (2021) [12] proposed the Spatial-Temporal Transformer Network (ST-TR), combining a Graph Convolutional Network (GCN) and a transformer to enhance human activity recognition (HAR) by capturing global attention across spatial and temporal dimensions. The model achieved a cross-subject accuracy of 88.60% and a cross-view accuracy of 94.70%, effectively addressing spatial and temporal complexities in HAR. Ref. [13] introduced the Temporal-Aware Adaptive Skeleton Graph Network (TA-ASGN), which combines transformers with temporal adaptive skeleton graphs for improved HAR. The model achieved a cross-subject accuracy of 89.80% and a cross-view accuracy of 95.30%, excelling in subject variation and viewpoint adaptation. Duan et al. (2022) [14] developed a Graph Convolution and Transformer Hybrid Model for skeleton-based action recognition. By integrating GCNs for local spatial feature extraction and transformers for global temporal attention, the model achieved a cross-subject accuracy of 90.10% and a cross-view accuracy of 96.20%, demonstrating robustness in spatio-temporal action recognition. Zhao et al. (2022) [15] introduced Skeleton-Aware Geometry Feature Learning for HAR, which leverages geometric relationships in skeleton data to improve recognition accuracy. This method achieved a cross-subject accuracy of 90.00% and a cross-view accuracy of 95.80%, refining skeleton-based HAR through geometry-aware feature learning.

Ref. [16] introduced Temporal Edge Aggregation for GCN enhancing HAR by aggregating temporal edge information. This method achieved a cross-subject accuracy of 90.30% and a cross-view accuracy of 96.00%, improving the temporal modeling capabilities of GCN-based HAR systems. Ref. [17] proposed a hybrid model combining a GCN with a transformer to enhance skeleton-based HAR. By incorporating an attention mechanism, the model selectively focuses on important spatial-temporal features, achieving a cross-subject accuracy of 89.70% and a cross-view accuracy of 95.70%, demonstrating its effectiveness in capturing local and global dependencies. Ref. [18] introduced the Dual Stream Transformer GCN, a model integrating both temporal and spatial streams to capture dynamic temporal changes and spatial relationships. The model achieved a cross-subject accuracy of 91.00% and a cross-view accuracy of 96.50%, highlighting its superior performance in HAR tasks through effective combination of temporal and spatial features. Ref. [11] achieved over 94% accuracy on URFD and UPFD datasets using Graph Convolutional Networks (GCNs) combined with Convolutional Neural Networks (CNNs).

Moreover, existing systems are rarely evaluated on vision-based datasets designed explicitly for MRHA recognition beyond falls. Two promising datasets, NTU RGB+D 120 and HMDB51, include medical-related human activity video data, offering new opportunities for MRHA recognition research. These datasets encompass a variety of medical activities, such as sneezing, coughing, sitting, and walking, which are relevant to healthcare scenarios. Despite this, few researchers have developed MRHA recognition systems based on these datasets, and their reported accuracy levels remain unsatisfactory. Addressing the limitations of existing approaches and leveraging these datasets for MRHA recognition could lead to the development of more effective, privacy-preserving, and versatile systems. Future work should focus on advancing vision-based

MRHA recognition models, improving their accuracy and generalizability across diverse populations and activity types.

3 Dataset Description

This study leverages existing HMR datasets, prioritizing datasets based on three criteria: (1) inclusion of activities listed in Table 1, (2) relevance to the medical domain, and (3) richness of features in the video dataset. Among the analyzed datasets, NTU RGB+D 120 emerged as the most suitable due to its overlap with key activities and its potential for improvement, as highlighted in the literature https://rose1.ntu.edu.sg/dataset/actionRecognition/ (accessed on 28 April 2025).

Table 1: Human Motion Recognition (HMR) datasets with some class activities

S.N.	Dataset name	Total samples data	Total classes
1	ActivityNet	21,313	200
2	Charades	66,493	157
3	HMDB51	6766	51
4	NTU RGB+D 120	114,480	120
5	STAIR Actions	109,478	100
6	UCF101	13,320	101

3.1 "NTU RGB+D 120" Dataset

The NTU RGB+D 120 dataset, developed by Nanyang Technological University, is a benchmark for human action recognition, featuring 120 motion classes and 114,480 samples captured in RGB video format (.mp4) at 1920 × 1080 resolution with 24 fps. Each sample includes 3D coordinates of 25 body joints, enabling detailed skeletal motion analysis. For this study, 12 healthcare-related activities were selected, such as sneezing/cough, staggering, and chest pain, totaling 13,200 samples, divided into 80% training (10,560 samples) and 20% testing (2640 samples) splits.

3.2 "HMDB51" Dataset

We conducted experiments on the HMDB51 dataset, a benchmark for human action recognition. The HMDB51 dataset comprises 6766 video clips with a total file size of 2 GB. It features 51 action categories, with each category containing at least 101 video clips sourced from diverse origins, including movies, YouTube, and other online platforms. For this study, we selected six classes: walk, stand, eat, sit, and drink—particularly relevant to medical and healthcare applications. These classes were chosen due to their critical importance in healthcare scenarios where activity recognition can provide meaningful insights and enhance patient monitoring systems.

4 Proposed Method

There are many researchers who have been working to develop a human motion recognition (HMR) system. However, a few researchers have been working to develop IoT-integrating HMR systems. we propose a new HMR method that uses spatial and temporal features powered by multi-stage deep learning and integrated with IoT. First, we use EfficientNet to extract spatial features from skeleton frame sequences. EfficientNet is designed with seven Mobile Inverted Bottleneck Convolutions (MBConv) blocks. Each block includes a convolutional layer, a depthwise separable layer, and a squeeze-and-excitation (SE) module to

create optimized feature representations. These spatial features are then passed to ConvLSTM, which extracts spatio-temporal features, combining spatial and sequential information. The main components of our study are given below:

- **OpenPose Based BodyPose Extraction:** We employed OpenPose to extract 25 key points from the whole for each frame in the sequence.
- **Hybrid Deep Learning Architecture:** We introduce a novel hybrid deep learning model, ENConvLSTM, combining EfficientNet and ConvLSTM to address challenges in HMR, such as high computational demands, low accuracy, and adaptability. The proposed architecture consists of two key components: efficientnet to extract the spatial feature from input skeleton data. It utilizes seven Mobile Inverted Bottleneck Convolutions (MBConv) blocks. Each MBConv block comprises a 1×1 convolution layer for feature mapping and a depthwise separable convolution layer for dimensionality reduction. A squeeze-and-excitation (SE) module for adaptive feature recalibration. Then we extracted the spatio-temporal feature within ConvLSTM. This takes the spatial features generated by EfficientNet and models the temporal dependencies across frames, producing robust spatio-temporal features.
- **Classification Module:** The spatio-temporal features generated by ConvLSTM are passed through a classification module, which includes Global Average Pooling and Fully Connected Layer, etc. This multi-stage pipeline ensures robust motion recognition by integrating spatial and temporal modeling techniques.
- **IoT Integration for Real-Time Alerts:** The system integrates IoT components, including a Raspberry Pi and GSM module, to provide real-time alerts. Twilio's SMS API is used to send instant notifications to caregivers and patients, removing the dependency on third-party mobile applications. This feature enhances the system's scalability and usability for healthcare scenarios.

4.1 Data Preprocessing

The preprocessing and pose keypoint extraction process begins with detecting 25 skeletal joints of the human body using the Kinect v2 camera. These joints include key points such as the head, shoulders, elbows, wrists, hips, knees, ankles, and feet, each represented by 3D coordinates (X, Y, Z). RGB videos are recorded at a resolution of 1920×1080, while depth maps are captured at 512×424 resolution. Skeletons are extracted from video frames at a rate of 24 frames per second (FPS), enabling precise tracking of motion trajectories. Fig. 2 illustrates the skeletal configuration, which abstracts human poses and movements while preserving spatial and temporal features crucial for motion analysis. The OpenPose library is used to extract skeletal data, identifying and tracking the 25 body joints for each individual in the scene. Each frame provides a reduced yet comprehensive skeletal representation of human movement, significantly simplifying raw video data. This abstraction captures essential motion patterns, allowing for the analysis of complex motion dynamics while reducing computational complexity. The skeleton data structure provides an efficient input format for deep learning models, focusing on critical movement patterns. The preprocessing pipeline further enhances the data for analysis. Video frames are sampled at 10 frames per second (FPS) to eliminate redundancy while retaining critical motion information. Each frame is resized to a uniform resolution, converted to grayscale, and normalized to ensure consistency and compatibility across samples. Finally, the processed frames are organized into sequential arrays to represent the temporal dynamics of motion and fed into the feature extraction module.

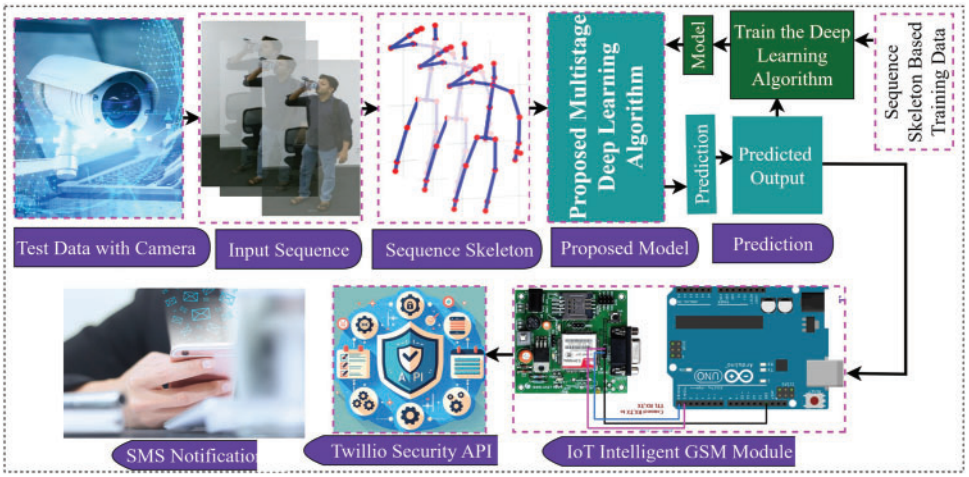

Figure 2: The workflow of the proposed methodology this figure represents an end-to-end system for human motion detection using a deep learning model, followed by real-time monitoring and notification

4.2 Spatial Temporal Feature Extraction

The sequential skeleton information is fed into the feature extraction module. Here first we extract the spatial feature using EfficientNet and then we fed him into ConvLSTM to extract the spatiotemporal features defined below.

4.2.1 EfficientNet Model

EfficientNet [19] is a state-of-the-art deep learning model for spatial feature enhancement and image classification tasks. It achieves high accuracy with fewer parameters and lower computational costs by scaling depth, width, and resolution in a balanced manner due to the depthwise separable convolution and sequence excitation module. Fig. 3b shows the efficient net model diagram, which was constructed with various deep learning modules to extract the spatial feature from input skeleton data. It utilizes seven Mobile Inverted Bottleneck Convolutions (MBConv) blocks. Each MBConv block comprises a 1×1 convolution layer for feature mapping and a depthwise separable convolution layer for dimensionality reduction. A squeeze-and-excitation (SE) module for adaptive feature recalibration. The series of MBConv is mainly used to downsample and extract meaningful features from the input, which is demonstrated in Fig. 3d. The structure consists of multiple blocks of convolutions, where the operations can be represented as:

$$y = f(x; W) = \text{MBConv}(x; W) \tag{1}$$

where x is the input skeleton features, W are the convolutional weights, and y is the output feature map. EfficientNet progressively reduces the spatial resolution while expanding the depth of features, leading to high-level representations for the next stage. These blocks apply depth-wise separable convolutions to efficiently extract hierarchical spatial features, resulting in feature maps from progressively lower resolutions and deeper feature representations.

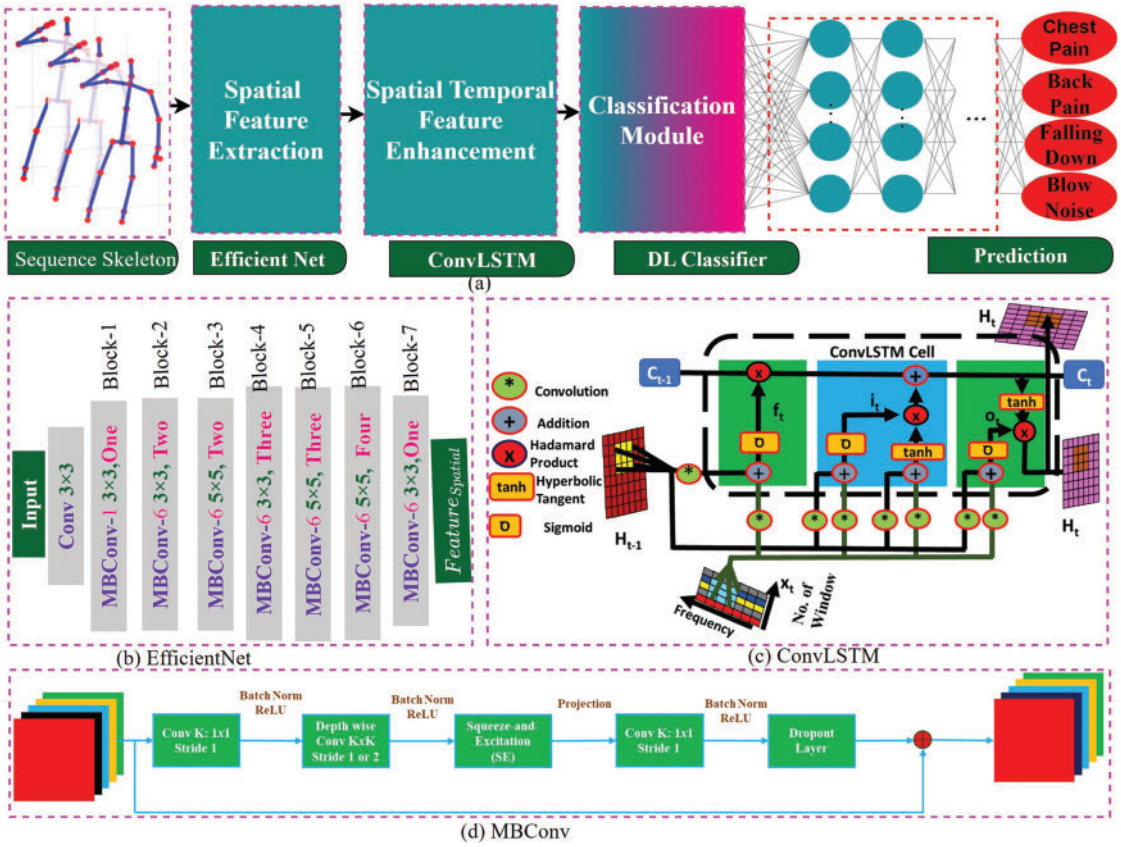

Figure 3: **(a)** Proposed multi-stage deep learning model constructed with **(b)** EfficientNet and **(c)** ConvLSTM beside the classification module **(d)** Mobile Inverted Bottleneck Convolutions [20]

4.2.2 ConvLSTM Model

Then we extracted the spatio-temporal feature within ConvLSTM [21]. This takes the spatial features generated by EfficientNet and models the temporal dependencies across frames, producing robust spatio-temporal features. This hybrid approach addresses the limitations of existing methods, such as high computational complexity, low accuracy, and limited adaptability to diverse healthcare scenarios. Fig. 3c demonstrated the ConvLSTM network diagram, which mainly extended the capabilities of traditional LSTM networks by integrating convolutional layers.

4.2.3 ENConvLSTM Proposed Model Architecture

Fig. 3 presents a hybrid architecture for human motion recognition, combining EfficientNet for spatial feature extraction and ConvLSTM for temporal feature modeling. At the top, the EfficientNet architecture is shown, which processes input frames (224 × 224 resolution) through a series of MBConv blocks. Once skeleton data is extracted, it is passed through the EfficientNet model for spatial feature extraction.

5 Experiment and Result

The proposed model is evaluated on the NTU RGB+D 120 dataset, focusing on 12 selected medical classes. The dataset is divided into training (80%) and testing (20%) portions. The model is trained on 80% of the data, with its performance evaluated on the remaining 20%. During training, the ENConvLSTM model

minimizes classification loss, typically cross-entropy, using the Adam optimizer to learn and differentiate between various human activities. Key performance metrics accuracy, precision, recall, and F1-score are calculated for both cross-subject and cross-view evaluations [22–24]. The model performs exceptionally well, particularly in cross-view evaluations, which involve more significant variability due to different camera angles, demonstrating its robustness and generalizability for real-time Human Motion Recognition (HMR) in healthcare applications.

5.1 Software and Hardware Requirements

The study requires several key software and hardware components. The software includes deep learning frameworks like TensorFlow, PyTorch, scikit-learn, and Keras, with data processing libraries such as NumPy, Pandas, and OpenCV, all using Python. Development is done in Jupyter Notebook and PyCharm. The hardware setup features an AMD Ryzen 9 5900X 12-Core Processor, running on a 64-bit system with Python 3.9.13 and CUDA 11.0. It includes an NVIDIA® GeForce RTX 3060 graphics card with 6 GB of memory, 64 GB of RAM, and a 4 TB SSD for storage. An Arduino UNO REV3 and a SIM900A GSM module are also used for SMS communication.

5.2 Ablation Study

The ablation study, as shown in Table 2, compares the performance of the proposed model with several baseline methods on the HMDB51 dataset using accuracy, precision, recall, and F1-score. It shows that the proposed combination achieves high performance accuracy compared to the baseline individual methods.

Table 2: Ablation study of the proposed model with HMDB51 dataset

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
Proposed (ENConvLSTM)	**89.22**	88.12	86.54	87.96

5.3 Proposed Model Parameters List

The proposed model has 8,138,104 parameters, of which 8,095,640 are trainable and 42,464 are non-trainable. The model is trained with a batch size of 16 for 100 epochs using sparse categorical cross-entropy as the loss function and Adam as the optimizer, with a learning rate of 0.001. Fig. 4 highlights Adam's superior performance over SGD, RMSprop, and Adagrad in training the proposed model. Adam achieves the highest validation accuracy (approaching 0.95) and the lowest validation loss due to its adaptive learning rate mechanism, which accelerates convergence and improves generalization. The smooth accuracy and loss curves further demonstrate Adam's stability and effectiveness, whereas other optimizers show slower convergence and higher validation losses. These results confirm Adam as the most effective optimizer for this model.

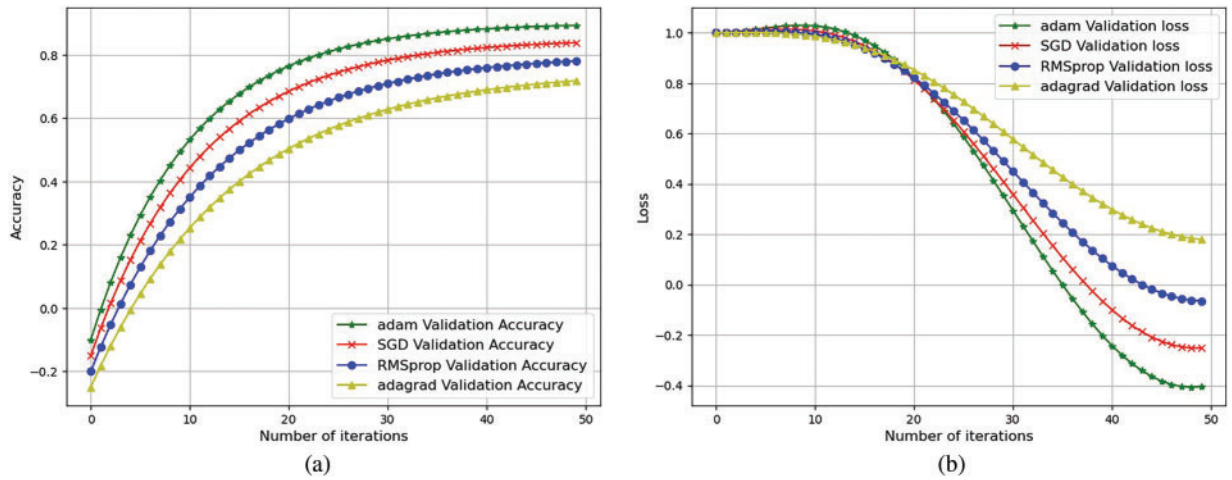

Figure 4: Model optimizers comparison: **(a)** Validation accuracy and **(b)** Validation loss curve for NTU dataset

5.4 Performance Matrix with NTU RGB+D 120 Dataset

We evaluated the proposed method with the NTU RGB+D medical related action dataset with various configurations, including cross-subject evaluation and cross-view evaluation. It involves splitting the dataset such that training and testing samples are from different subjects, ensuring the model generalizes well to unseen individuals. The model achieved a cross-subject accuracy of 94.85%. It involves splitting the dataset based on different camera views, ensuring the model performs well across various perspectives. The model achieved a cross-view accuracy of 96.45%. Key performance metrics—accuracy, precision, recall, and F1-score—are calculated for both cross-subject and cross-view evaluations, as shown in Table 3.

Table 3: Other performance evaluation metrics on NTU RGB+D 120 dataset

S.N.	Performance martics	Dataset evaluation methods	
		Cross-Subject (%)	Cross-View (%)
1	Accuracy	94.85	96.45
2	Precision	93.70	95.90
3	Recall	94.30	96.10
4	F1-Score	94.00	96.00

In the cross-subject evaluation, the accuracy exhibits a robust upward trend from an initial 75.01% to a final 96.85%, with some minor fluctuations and eventual stabilization. Concurrently, the loss decreases consistently from 0.60 to 0.04, reflecting an overall improvement in model performance. Similarly, the cross-view evaluation shows a strong accuracy progression, starting at 78.34% and reaching 96.45% by the end, with minor variations throughout and a stabilizing trend. The loss trend in this evaluation mirrors that of the cross-subject assessment, declining from 0.55 to 0.03, indicating effective learning and convergence. Both evaluations demonstrate an effective model performance enhancement over time, with accuracy improving and loss decreasing, culminating in stable performance metrics.

5.4.1 State of the Art Comparison for the NTU RGB+D 120 Dataset

The performance of the ENConvLSTM model is compared with several state-of-the-art models, including traditional methods and recent deep-learning approaches shows in Table 4. Each Performance Evaluation Method (Cross-View and Cross-Subject) is graphically presented in Fig. 5a,b. The proposed ENConvLSTM model significantly outperforms the other models across cross-subject and cross-view evaluations on the NTU RGB+D 120 dataset, showing excellent accuracy, precision, recall, and F1 score results. This indicates that the proposed model is highly robust and performs well across different subjects and viewing conditions.

Table 4: State of the art comparison for NTU RGB+D 120 dataset

Used model with citation	NTU RGB+D 120	
Methods	**Cross-Subject (%)**	**Cross-View (%)**
MS-G3D [25]	86.92	88.44
PA-ResGCN-B19 [26]	87.34	88.37
Dynamic GCN [27]	87.37	88.68
CTR-GCN [28]	88.99	90.62
4s Shift-GCN [29]	85.9	87.6
ST-TR [30]	89.95	96.12
GA-GCN [31]	92.38	92.83
ENConvLSTM (Proposed)	**94.85**	**96.45**

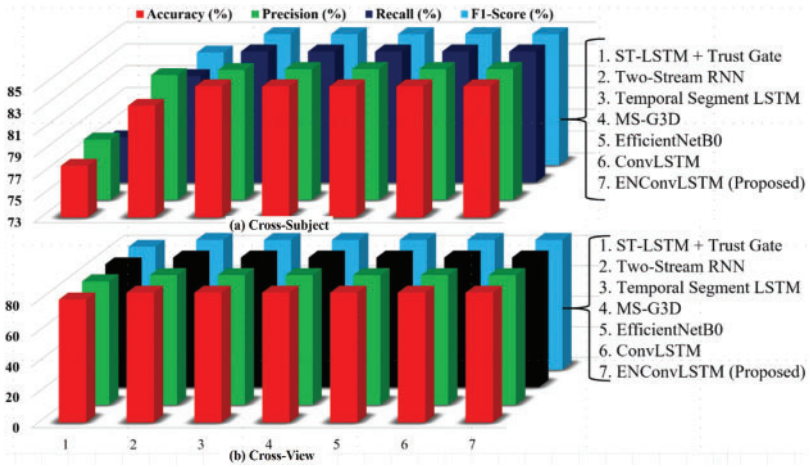

Figure 5: Performance comparison among different tested models on the NTU RGB+D 120 dataset with **(a)** cross-view **(b)** cross-subject configuration

5.5 Performance Accuracy and State of the Art Comparison for HMDB51 Dataset

Fig. 6 presents the confusion matrix and loss with various optimizations. Therefore, we consider the Adam optimizer for this study. This matrix offers a detailed view of accuracy for each individual class. Fig. 6a shows the confusion matrix of the proposed model for the HMDB15 dataset. From this matrix, we can see that the diagonal line represents the true classes, while the off-diagonal data represents false detections. The classes "drink," "eat," and "walk" achieve remarkable accuracy in this dataset. However, the other classes show comparatively lower accuracy. We analyze the validation loss curve relative to the number

of iterations, as illustrated in Fig. 6b. The curve indicates that the ADAM optimizer outperforms both the SGD and RMSProp algorithms. Based on the analysis, Adam is likely the best-performing optimizer among the three optimizers, such as Adam [32], SGD [33,34], and RMSProp [33,35]. This is because Adam combines the advantages of both RMSProp and momentum, leading to faster and more reliable convergence [33]. It adapts the learning rate during training and uses past gradient information to accelerate learning, making it efficient and effective for many deep learning problems [35]. These standard classification metrics for the HMDB51 dataset are listed in Table 5, where the average accuracy in our experiment is 89.22%. Table 6 provides a comparative analysis of various models tested on the HMDB51 dataset for human action recognition. The proposed model, a multi-stage model, achieves the highest accuracy of 89.22%, showcasing its superior performance. Models such as STM Framework [36] and Attention-Based LSTM with 3D CNN [37] demonstrate strong performance with accuracies of 80.40% and 87.98%, respectively, while EfficientNet delivers an impressive accuracy of 88.70% [38]. In contrast, earlier models like VicTR (B/16) [39] and SVT Self-Supervised Transfer [40] achieve relatively lower accuracies of 67.28% and 57.80%, respectively. Additionally, reference [41] proposed a Vision Transformer (ViT) model, achieving 59.74% and 68.2% accuracy on HMDB51 by leveraging self-attention. Moreover, reference [42] introduced a Dual-Stream Framework, achieving 78.62% accuracy by separately processing temporal and spatial features for improved action recognition, surpassing recent approaches over traditional methodologies.

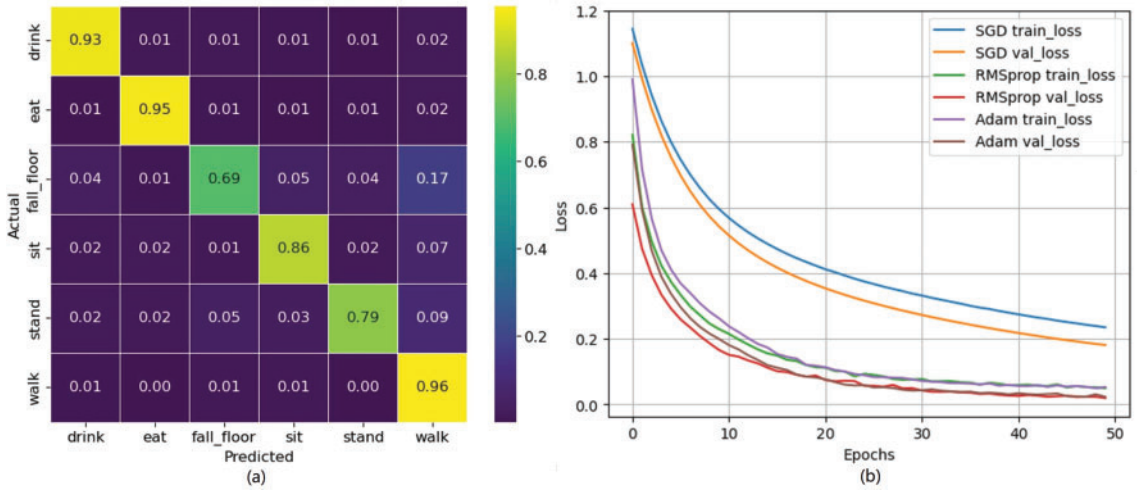

Figure 6: **(a)** Confusion matrix **(b)** Loss curve for the HMDB51 dataset

Table 5: Classification result for the HMDB51 dataset

Selected class levels	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
drink	89.00	93.00	91.00	–
eat	89.00	95.00	92.00	–
fall_floor	84.00	69.00	76.00	–
sit	85.00	86.00	86.00	–
stand	91.00	79.00	85.00	–
walk	91.00	96.00	93.00	–
Average	88.17	86.33	87.17	**89.22**

Table 6: State-of-the-art comparison for the proposed model with HMDB51 dataset (Sort by Year)

Author and citation	Model name	Year	Accuracy (%)
Ranasinghe et al. [40]	Self-supervised transfer	2022	67.28, 57.8
Sarraf et al. [41]	ViT	2023	59.74, 68.2
Saoudi et al. [37]	Attention-LSTM-3DCNN	2023	87.98
Burton-Barr et al. [38]	Act-control vision	2024	88.70
Kahatapitiya et al. [39]	VicTR (B/16)	2024	51.00
Hussain et al. [43]	Dual-stream framework	2024	78.62
Jiang et al. [36]	STM framework	2024	80.40
Proposed	**EfficientNetB0ConvLSTM**	**2024**	**89.22**

6 Real-Deployment

Finally, the proposed deep learning system is integrated with a Raspberry Pi and GSM module to send real-time alerts via Twilio SMS service, keeping caregivers and patients informed instantly. This system is scalable, efficient, and proactive, helping to improve patient monitoring and outcomes and reduce healthcare costs.

6.1 Real Time Implementation

Fig. 7 shows the interconnection between an Arduino UNO and the SIM900 GSM module to implement the real-time scenario. A 12V 2Amp DC adapter powers the Arduino. The TX (transmit) and RX (receive) pins are crucial for communication. The TX pin of the Arduino connects to the RXD pin of the TTL module for data transmission, while the RX pin of the Arduino connects to the RDX pin of the TTL module for receiving data. Both devices share a common ground (GND) for proper operation. This setup enables the Arduino to communicate with the GSM module for tasks like sending SMS or making voice calls. The connections include power and ground lines: the red wire represents the VCC (power) connection from the Raspberry Pi to the ESP8266, while the blue wire indicates the ground (GND) connection. The yellow wire connects the GPIO pin from the Raspberry Pi to the ESP8266's TX pin for data transmission, and the green wire connects another GPIO pin to the RX pin of the ESP8266 for data reception. This setup enables the Raspberry Pi to communicate with the ESP8266 for wireless connectivity in various projects, such as IoT applications. Arduino UNO REV3-Compact and Versatile Microcontroller Board with A000066. Interface a SIM900A GSM module with an Arduino to send and receive SMS. Arduino UNO REV3-Compact and Versatile Microcontroller Board with A000066. Interface a SIM900A GSM module with an Arduino to send and receive SMS. Fig. 3 illustrates the proposed system for human motion recognition using an ENConvLSTM model integrated with an Arduino. It begins with an input video that is processed to extract skeleton data, represented in a 3D coordinate system (X, Y, Z). This skeleton data is fed into the ENConvLSTM model, which then utilizes a SoftMax layer to predict the activity being performed. The predicted activity is communicated to an Arduino, which can be powered by a stable 5V source or battery, indicating a real-time or recorded data application. This setup enables effective monitoring and recognition of human activities through skeletal motion analysis. The proposed human motion recognition system's performance in the laboratory experiment is satisfactory. However, there are always some differences between laboratory and real-time scenarios. Table 7 illustrates the result in a real-time scenario, and its visual representation is shown in Fig. 8.

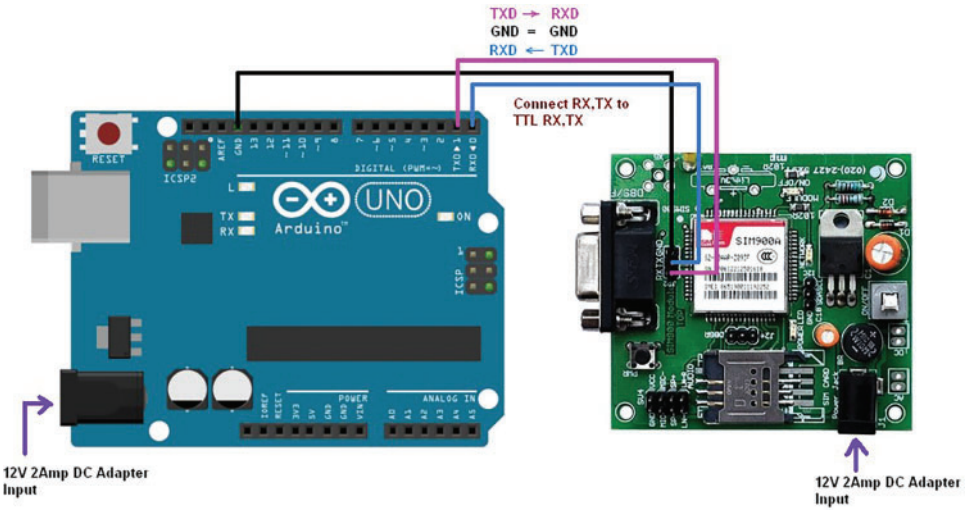

Figure 7: Schematic diagram illustrating the connections between an Arduino UNO and a SIM900 GSM TTL module, highlighting the power supply, TX and RX pin connections, and the common ground to facilitate communication for tasks such as SMS sending and voice calling

Table 7: Laboratory and real-time experiment performance for the HMDB51 dataset

No.	Actual class	Real-time observation (individual)	Laboratory experiment (%)	Real-time experiment (%)	Calculate score (%)	Loss of real-time experiment (%)
1	fall_floor	Individual fall floor	89.00	87.00	97.75	2.00
2	walk	Individual walking	95.00	91.00	95.79	4.00
3	stand	Individual standing	79.00	77.00	97.47	2.00
4	eat	Individual eating	94.00	90.00	95.74	4.00
5	sit	Individual sitting	86.00	82.00	95.35	4.00
6	drink	Individual drinking	92.00	88.00	95.65	4.00
Average score			**89.22**	**85.83**	**96.29**	**3.33**

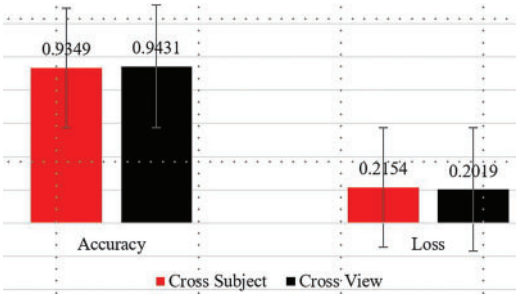

Figure 8: Real-Time human motion recognition performance metrics (Includes Accuracy and Loss Values) [NTU RGB+D 120 dataset]

6.2 Sending Alert Message (Notification)

In our proposed system, the model predicts human actions, specifically identifying whether an individual is experiencing a fall. Upon detecting a fall event, the system promptly triggers an alert message to the registered user mobile device. This real-time communication is facilitated through the integration of Twilio [44], a leading SMS and communication server provider, ensuring reliable and instantaneous notifications. By leveraging Twilio's robust platform, we enhance the responsiveness of our human motion recognition system, thereby significantly improving user safety and emergency response efficiency.

6.3 Work Limitations

While our paper highlights the integration of IoT components like Raspberry Pi and the GSM module, we acknowledge limitations in latency, reliability, and scalability in real-time scenarios. The system is not optimized for handling multiple alerts through efficient data transmission and prioritization, which could be explored in future work. Moreover, challenges such as imbalanced class distributions and noise in real-world data may also affect model performance. Future research could explore differential privacy and federated learning for enhanced security, particularly in the context of sensitive patient data.

7 Conclusion and Future Work

This study presents a novel IoT-based framework for real-time Medical-Related Human Activity (MRHA) recognition, addressing challenges like high computational demands and low accuracy in existing systems. By combining EfficientNet for spatial feature extraction and ConvLSTM for spatio-temporal integration, the method achieves strong performance, with 94.85% accuracy for cross-subject and 96.45% for cross-view evaluations on the NTU RGB+D 120 dataset. It also demonstrates 89.00% accuracy on the HMDB51 dataset. A key contribution is integrating the MRHA system with a Raspberry Pi and GSM module, providing real-time alerts through SMS. The system shows promise for improving patient monitoring and healthcare outcomes, although it may face challenges in real-time applications due to environmental factors. Future work will focus on multimodal datasets, cloud computing for remote monitoring, and addressing privacy concerns, ensuring the system practical use in healthcare settings.

Acknowledgement: I would like to express my sincere gratitude to the ICT Division of the Ministry of Posts, Telecommunications, and Information Technology of the People's Republic of Bangladesh for their invaluable support.

Funding Statement: This research was funded by the ICT Division of the Ministry of Posts, Telecommunications, and Information Technology of Bangladesh under Grant Number 56.00.0000.052.33.005.21-7 (Tracking No. 22FS15306), with support from the University of Rajshahi.

Author Contributions: Md. Ekramul Hamid led the research and supervised the project. Subrata Kumer Paul and Abu Saleh Musa Miah designed the methodology; Subrata Kumer Paul also collected and processed data. Software was implemented by Rakhi Rani Paul, Subrata Kumer Paul, and Md. Ekramul Hamid. The draft was written by Subrata Kumer Paul, Abu Saleh Musa Miah, and Rakhi Rani Paul, with all authors contributing to revisions. Jungpil Shin handled administration and funding, while Md Abdur Rahim oversaw validation and review. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: NTU RGB+D 120 Dataset: https://rose1.ntu.edu.sg/dataset/actionRecognition/ (accessed on 28 April 2025); Github Source Code: www.github.com/Subrata11/IoT-based-Real-time-Human-Motion-Recognition-Based-on-Skeletons- (accessed on 28 April 2025).

Ethics Approval: Not applicable.

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

References

1. Lord SR. Visual risk factors for falls in older people. Age Ageing. 2006;35(Suppl 2):ii42–5. doi:10.1093/ageing/afl085.
2. Egawa R, Miah ASM, Hirooka K, Tomioka Y, Shin J. Dynamic fall detection using graph-based spatial temporal convolution and attention network. Electronics. 2023;12(15):3234. doi:10.3390/electronics12153234.
3. Hassan N, Miah ASM, Shin J. A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. Appl Sci. 2024;14(2):603. doi:10.3390/app14020603.
4. Hassan N, Miah ASM, Shin J. Enhancing human action recognition in videos through dense-level features extraction and optimized long short-term memory. In: 2024 7th International Conference on Electronics, Communications, and Control Engineering (ICECC); 2024 Mar 22–24; Kuala Lumpur, Malaysia. IEEE; 2024. p. 19–23. doi:10.1109/ICECC63398.2024.00011.
5. Romeo L, Marani R, Petitti A, Milella A, D'Orazio T, Cicirelli G. Image-based mobility assessment in elderly people from low-cost systems of cameras: a skeletal dataset for experimental evaluations. In: Ad-hoc, mobile, and wireless networks. Cham. Springer; 2020. p. 125–30. doi:10.1007/978-3-030-61746-2_10.
6. Paul SK, Miah ASM, Paul RR, Hamid ME, Shin J, Rahim MA. IoT-based real-time medical-related human activity recognition using skeletons and multi-stage deep learning for healthcare. arXiv:2501.07039. 2025.
7. Gutiérrez J, Rodríguez V, Martin S. Comprehensive review of vision-based fall detection systems. Sensors. 2021;21(3):947. doi:10.3390/s21030947.
8. Huang Z, Liu Y, Fang Y, Horn BKP. Video-based fall detection for seniors with human pose estimation. In: 2018 4th International Conference on Universal Village (UV); 2018 Oct 21–24; Boston, MA, USA. IEEE; 2018. p. 1–4. doi:10.1109/UV.2018.8642130.
9. Yu X, Wang C, Wu W, Xiong S. A real-time skeleton-based fall detection algorithm based on temporal convolutional networks and transformer encoder. Pervasive Mob Comput. 2025;107:102016. doi:10.1016/j.pmcj.2025.102016.
10. Akash HS, Rahim MA, Miah ASM, Lee HS, Jang SW, Shin J. Two-stream modality-based deep learning approach for enhanced two-person human interaction recognition in videos. Sensors. 2024;24(21):7077. doi:10.3390/s24217077.
11. Zahan S, Hassan GM, Mian A. SDFA: structure-aware discriminative feature aggregation for efficient human fall detection in video. IEEE Trans Ind Inform. 2023;19(8):8713–21. doi:10.1109/TII.2022.3221208.
12. Plizzari C, Cannici M, Matteucci M. Spatial temporal transformer network for skeleton-based action recognition. In: ICPR International Workshops and Challenges. Cham. Springer; 2021.
13. Shi L, Zhang Y, Cheng J, Lu H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In: Proceedings of the IEEE/CVP Conferenceon Computer Vision and Pattern Recognition (CVPR); 2019. p. 12026–35.
14. Duan H, Wang J, Chen K, Lin D. DG-STGCN: dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv:2210.05895.2022.

15. Usama MM, Ali H, Ahmed J, Ashraf R, Mahmood S Ahmad J. Learning skeleton aware geometry features for action recognition. ST-RTR: spatial temporal relative transformer for skeleton-based human action recognition. arXiv:2410.23806. 2024.
16. Heidari N, Iosifidis A. Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy; 2021. p. 7907–14. doi:10.1109/ICPR48806.2021.9412091.
17. Yang W, Zhang J, Cai J. HybridNet: integrating GCN and CNN for skeleton-based action recognition. Appl Intell. 2023;53:574–85. doi:10.1007/s10489-022-03436-0.
18. Chen D, Chen M, Wu P. Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. Sci Rep. 2024;15:4982. doi:10.1038/s41598-025-87752-8
19. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946. 2019.
20. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. IEEE; 2018. p. 4510–20. doi:10.1109/CVPR.2018.00474.
21. Tang J, Zhu R, Wu F, He X, Huang J, Zhou X, et al. Deep spatio-temporal dependent convolutional LSTM network for traffic flow prediction. Sci Rep. 2025;15(1):11743. doi:10.1038/s41598-025-95711-6.
22. Shin J, Miah ASM, Egawa R, Hassan N, Hirooka K, Tomioka Y. Multimodal fall detection using spatial-temporal attention and Bi-LSTM-based feature fusion. Future Internet. 2025;17(4):173. doi:10.3390/fi17040173.
23. Miah ASM, Al Mehedi Hasan M, Jang SW, Lee HS, Shin J. Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition. Electronics. 2023;12(13):2841. doi:10.3390/electronics12132841.
24. Miah ASM, Al Mehedi Hasan M, Shin J. Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model. IEEE Access. 2023;11:4703–16. doi:10.1109/access.2023.3235368.
25. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. IEEE; 2020. p. 140–9. doi:10.1109/cvpr42600.2020.00022.
26. Song YF, Zhang Z, Shan C, Wang L. Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020; Seattle, WA, USA. ACM. p. 1625–33. doi:10.1145/3394171.3413802.
27. Ye F, Pu S, Zhong Q, Li C, Xie D, Tang H. Dynamic GCN: context-enriched topology learning for skeleton-based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020; Seattle, WA, USA. ACM. p. 55–63. doi:10.1145/3394171.3413941.
28. Chen Y, Zhang Z, Yuan C, Li B, Deng Y, Hu W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. IEEE; 2021. p. 13339–48. doi:10.1109/ICCV48922.2021.01311.
29. Cheng K, Zhang Y, He X, Chen W, Lu H. Skeleton-based action recognition with shift graph convolutional network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Seattle, WA, USA, 2020. pp. 180–9. doi:10.1109/CVPR42600.2020.00026.
30. Plizzari C, Cannici M, Matteucci M. Skeleton-based action recognition via spatial and temporal transformer networks. Comput Vis Image Underst. 2021;208(3):103219. doi:10.1016/j.cviu.2021.103219.
31. Abduljalil H, Elhayek A, Marish Ali A, Alsolami F. Spatiotemporal graph autoencoder network for skeleton-based human action recognition. AI. 2024;5(3):1695–708. doi:10.3390/ai5030083.
32. Kumer Paul S, Ala Walid MA, Rani Paul R, Uddin MJ, Rana MS, Kumar Devnath M, et al. An Adam based CNN and LSTM approach for sign language recognition in real time for deaf people. Bulletin EEI. 2024;13(1):499–509. doi:10.11591/eei.v13i1.6059.
33. Mulyono IUW, Kusumawati Y, Susanto A, Sari CA, Islam HMM, Doheir M. Hiragana character classification using convolutional neural networks methods based on Adam, SGD, and RMSProps optimizer. Sci J Informatics. 2024;11(2):467–76. doi:10.15294/sji.v11i2.2313.
34. Merrouchi M, Atifi K, Skittou M, Benyoussef Y, Gadi T. AutoLrOpt: an efficient optimizer using automatic setting of learning rate for deep neural networks. IEEE Access. 2024;12(8):83154–68. doi:10.1109/access.2024.3413043.

35. Ahda FA, Wibawa AP, Dwi Prasetya D, Arbian Sulistyo D. Comparison of Adam optimization and RMS prop in minangkabau-Indonesian bidirectional translation with neural machine translation. JOIV Int J Inform Vis. 2024;8(1):231. doi:10.62527/joiv.8.1.1818.
36. Jiang B, Wang M, Gan W, Wu W, Yan J. STM: spatiotemporal and motion encoding for action recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. IEEE; 2019. p. 2000–9. doi:10.1109/ICCV.2019.00209.
37. Saoudi EM, Jaafari J, Andaloussi SJ. Advancing human action recognition: a hybrid approach using attention-based LSTM and 3D CNN. Sci Afr. 2023;21(3):e01796. doi:10.1016/j.sciaf.2023.e01796.
38. Burton-Barr J, Fernando B, Rajan D. Activation control of vision models for sustainable AI systems. IEEE Trans Artif Intell. 2024;5(7):3470–81. doi:10.1109/TAI.2024.3372935.
39. Kahatapitiya K, Arnab A, Nagrani A, Ryoo MS. VicTR: video-conditioned text representations for activity recognition. arXiv:2304.02560. 2023.
40. Ranasinghe K, Naseer M, Khan S, Khan FS, Ryoo M. Self-supervised video transformer. arXiv:2112.01514. 2021.
41. Sarraf S, Kabia M. Optimal topology of vision transformer for real-time video action recognition in an end-to-end cloud solution. Mach Learn Knowl Extr. 2023;5(4):1320–39. doi:10.3390/make5040067.
42. Hussain A, Hussain T, Ullah W, Baik SW. Vision transformer and deep sequence learning for human activity recognition in surveillance videos. Comput Intell Neurosci. 2022;2022(3):3454167. doi:10.1155/2022/3454167.
43. Hussain A, Khan SU, Khan N, Ullah W, Alkhayyat A, Alharbi M, et al. Shots segmentation-based optimized dual-stream framework for robust human activity recognition in surveillance video. Alex Eng J. 2024;91(9):632–47. doi:10.1016/j.aej.2023.11.017.
44. Paul SK, Zisa AA, Ala Walid MA, Zeem Y, Paul RR, Haque MM, et al. Human fall detection system using long-term recurrent convolutional networks for next-generation healthcare: a study of human motion recognition. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2023 Jul 6–8; Delhi, India. IEEE; 2023. p. 1–7. doi:10.1109/ICCCNT56998.2023.10308247.