International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-12 Issue-3, February 2023

IoT-Based Real-Time Human Motion Recognition Based on Skeletons for Next Generation Healthcare using Deep Learning Techniques

Subrata Kumer Paul, Rakhi Rani Paul, Md. Ekramul Hamid

Abstract*: The Internet of Things (IoT) and mobile technology have transformed healthcare by enabling real-time monitoring and diagnosis of patients. Human Motion Recognition (HMR) plays a vital role in healthcare systems in the 21st century, yet existing solutions face challenges such as high computational demands, low accuracy, and limited adaptability. This paper introduces a novel approach to real-time HMR using skeleton data with ENConvLSTM. This hybrid deep learning model combines EfficientNet and ConvLSTM to improve both accuracy and processing speed. The EfficientNet excels in image classification, while ConvLSTM processes sequential data, making this hybridization well-suited for dynamic motion analysis. Applied to the "NTU RGB+D 120" public dataset, our model achieves an accuracy of 94.85% for cross-subject and 96.45% for cross-view evaluations, addressing key healthcare-related activities across 12 medical classes. Moreover, the dataset presents several challenges, including a large data size, high dimensionality, class imbalance, noisy data, and limited domain labelling, all of which require advanced processing and modelling techniques. Our model's unique use of skeleton data tracking 25 body joints enables the detection of subtle motion changes, ideal for early diagnosis and intervention in healthcare scenarios. Additionally, we integrate the system with a Raspberry Pi and a GSM module, providing real-time alerts via Twilio's SMS service to instantly notify caregivers and patients. However, this solution is pivotal in modern healthcare, offering scalable, efficient, and proactive patient monitoring, leading to improved outcomes and reduced healthcare costs.

***Keywords*: Real-time Human Motion Recognition (HMR), EN Conv LSTM, Skeleton Data, NTU RGB+D 120 Dataset**

I. INTRODUCTION

Human Motion Recognition (HMR) systems aim to automatically identify and monitor individual or group activities, a focus that has significant implications for healthcare, particularly in tracking physical activity. Accurate detection remains a critical challenge in this field. With the global elderly population projected to reach 2.1 billion by 2050, according to the WHO and the United Nations, there is a growing strain on nursing home services [1]. Currently, many elderly individuals reside alone in care facilities where the number of patients rapidly exceeds the number of available healthcare professionals, leading to inadequate monitoring. This deficiency in surveillance can result in serious incidents, including falls, which the WHO reports cause approximately 646,000 deaths and over 37 million serious injuries annually. Addressing these issues through improved human motion recognition could enhance safety and care for the ageing population. The integration of the Internet of Things (IoT) and mobile technology into the healthcare sector has significantly transformed the way patient monitoring and diagnosis are conducted. This enables real-time, continuous oversight of patients' health conditions [2].

Focusing on skeleton data, which captures comprehensive joint movements, this research aims to provide a more effective and robust solution for real-time HMR in healthcare, ultimately contributing to improved patient outcomes and reduced healthcare costs [3]. The primary objective of this research is to apply the ENConvLSTM technique to develop a real-time motion recognition system integrated with IoT infrastructure. By implementing a mobile phone notification system that operates independently of third-party mobile apps, the research aims to ensure seamless communication of critical health-related events.

The accuracy and effectiveness of this system will be rigorously validated, contributing to innovative advancements in healthcare by providing timely, data-driven insights for improved patient outcomes. Our main contributions are summarized as follows:

☑ Dataset selection for skeleton-based research on the next generation Health care system.

☑ Focus on 12 relevant medical classes (e.g., sneezing, falling, chest pain), utilizing skeleton data to detect Subtle motion changes for early diagnosis and intervention.

☑ Apply the model successfully to the large size and complex "NTU RGB+D 120" dataset, using techniques like data augmentation and balanced sampling to address challenges such as class imbalance and noisy data.

☑ Designing the EN Conv LSTM model, which is a combination of Efficient Net and Conv LSTM, to handle both spatial and temporal features in skeleton-based HMR effectively.

☑ Enable real-time motion

Manuscript received on 08 January 2023 | Revised Manuscript received on 29 January 2023 | Manuscript Accepted on 15 February 2023 | Manuscript published on 28 February 2023.
*Correspondence Author(s)
Subrata Kumer Paul*, Department of CSE, University of Rajshahi, Rajshahi, Bangladesh, Department of CSE, Bangladesh Army University of Engineering & Technology (BAUET), Natore, Rajshahi, Bangladesh. Email ID: sksubrata96@gmail.com, ORCID ID: 0000-0001-9427-8350
Rakhi Rani Paul, Department of CSE, University of Rajshahi, Rajshahi, Bangladesh. Department of CSE, Bangladesh Army University of Engineering & Technology (BAUET), Natore, Rajshahi, Bangladesh. Email ID: rakhipaul.cse@gmail.com, ORCID ID: 0000-0001-7037-5850
Md. Ekramul Hamid, Department of CSE, University of Rajshahi, Rajshahi, Bangladesh. Email ID: ekram_hamid@yahoo.com, ORCID ID: 0000- 0001-8029-013X.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

recognition and notifications by integrating the model with a Raspberry Pi and GSM module, which is suitable for healthcare applications.

☑ Advance the field of human motion recognition by providing a robust, efficient solution addressing both spatial and temporal data aspects, with significant potential for improving patient care and reducing healthcare costs.

☑ Integrate a cloud-based mobile notification system that enables swift responses to medical emergencies by sending SMS to connected people.

☑ No need for 3rd mobile application. SMS notification is sent through the GSM Module and uses the "Twilio" SMS API., using its web service API.

The research significantly enhances healthcare by providing real-time monitoring for vulnerable populations, including the elderly and those with chronic illnesses, enabling early diagnosis and timely intervention in medical emergencies. It reduces caregiver burden, lowers healthcare costs, and expands access to healthcare in remote and underserved areas. By fostering patient empowerment and supporting ageing in place, the system promotes independence while reducing the incidence of preventable injuries. The use of IoT and cloud-based notifications ensures quick responses, helping to build a future of connected, proactive healthcare.

The remainder of the paper is organised into the following sections. The second section contains the Literature Review. The significant contributions are included in Section Two. The dataset description, its selection criteria, and the preprocessing steps are mentioned in Section Four. The fifth section presents the proposed methodology. The experimental results and performance evaluation are presented in Section 6 of this paper. The real-time implementation and its analysis are presented in Section 7. We have discussed the limitations and future scope of the proposed system in the eight sections. Finally, the paper is concluded, and the Future direction of this research is presented in Section 9. The remaining parts include acknowledgement, author contributions, conflicts of interest, data availability statement, list of abbreviations, and references.

II. 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 has been explored through various methodologies over the past few years. In this survey, we consider only the "NTU RGB+D 120" dataset. Cross-Subject (CS) and Cross-View (CV) are two methods for evaluating accuracy. The following paragraph presents the existing works with achieved accuracy.

Peng et al. [4] proposed a novel architecture for human activity recognition (HAR) that integrates a transformer model with a Progressive Spatio-Temporal Graph Convolutional Network (PST-GCN). This approach effectively captures both spatial and temporal dependencies within activity sequences, combining the strengths of both GCN and transformer techniques. By applying this architecture, the model achieved impressive results, with a cross-subject accuracy of 89.00% and a cross-view accuracy of 95.00%. These results highlight the model's robustness in handling both subject variations and changes in viewpoint, which are critical challenges in HAR systems.

Chen et al. [5] introduced a Hierarchical Aggregation Graph Convolutional Network (GCN) designed to improve human activity recognition (HAR) by effectively combining local and global action representations. This hierarchical structure enhances the model's ability to capture complex interactions between different body parts and activities. The approach demonstrated strong performance, achieving 87.50% cross-subject accuracy and 93.90% cross-view accuracy, underscoring its effectiveness in generalizing across diverse subjects and viewpoints in HAR tasks.

Cai et al. [6] developed the Joint-Bone Fusion Graph Convolutional Network (J-BGCN), which focuses on enhancing human activity recognition (HAR) by modelling both joint and bone relationships simultaneously. By combining these two key aspects of motion data, the network captures richer spatial information, resulting in improved recognition performance. The J-BGCN achieved 85.50% cross-subject accuracy and 92.60% cross-view accuracy, demonstrating its ability to effectively learn from joint-bone interactions and improve the overall robustness of HAR systems.

Song et al. [7] introduced the Adaptive Skeleton Graph Network (AdaSGN), which adaptively learns the structure of skeleton graphs to provide more flexible and dynamic representations for human activity recognition (HAR). This adaptive approach enables the model to capture the nuances of human movement more effectively, thereby enhancing its ability to generalise across various activities. AdaSGN achieved a cross-subject accuracy of 88.50% and a cross-view accuracy of 94.20%, demonstrating its effectiveness in handling diverse subject conditions and viewpoints in HAR tasks.

Plizzari et al. [8] proposed the Spatial-Temporal Transformer Network (ST-TR), which integrates a Graph Convolutional Network (GCN) with a transformer to enhance human activity recognition (HAR) by capturing global attention across spatial and temporal dimensions. This hybrid model leverages the GCN's ability to model local dependencies and the transformer's strength in capturing long-range interactions. The ST-TR achieved a cross-subject accuracy of 88.60% and a cross-view accuracy of 94.70%, showcasing its robustness in addressing both spatial and temporal complexities in HAR tasks.

Song et al. [9] introduced the Temporal-Aware Adaptive Skeleton Graph Network (TA-ASGN), a model that combines transformers with temporal adaptive skeleton graphs for improved human activity recognition (HAR). This approach enhances the flexibility of the skeleton graph while incorporating temporal dependencies. Through the transformer, allowing for more precise modelling of complex human movements. TA-ASGN achieved a cross-subject accuracy of 89.80% and a cross-view accuracy of 95.30%, demonstrating superior performance in both

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

subject variation and viewpoint adaptation in HAR tasks.

Duan et al. [10] developed a Graph Convolution and Transformer Hybrid Model designed for spatio-temporal skeleton-based action recognition in human activity recognition (HAR). By integrating the strengths of graph convolutional networks (GCNs) for local spatial feature extraction and transformers for global temporal attention, the model effectively captures complex activity patterns. This hybrid approach achieved impressive results, with a cross-subject accuracy of 90.10% and a cross-view accuracy of 96.20%, highlighting its robustness and effectiveness in handling both spatial and temporal aspects of skeleton-based action recognition.

Zhao et al. [11] introduced a novel approach to human activity recognition (HAR) through Skeleton-Aware Geometry Feature Learning, which leverages geometry-based skeleton features to enhance recognition accuracy. By focusing on the geometric relationships within the skeleton data, the model can capture more detailed movement patterns, leading to improved performance. The method achieved a cross-subject accuracy of 90.00% and a cross-view accuracy of 95.80%, demonstrating its effectiveness in refining skeleton-based HAR through geometry-aware feature learning.

Yang et al. [12] proposed an innovative approach to human activity recognition (HAR) by introducing Temporal Edge Aggregation for Graph Convolutional Networks (GCN). This technique focuses on enhancing the recognition of skeleton data by aggregating temporal edge information, which improves the model's ability to capture dynamic and sequential aspects of human movements. The method achieved a cross-subject accuracy of 90.30% and a cross-view accuracy of 96.00%, highlighting its effectiveness in enhancing the temporal modelling and recognition capabilities of GCN-based HAR systems.

Feng et al. [13] introduced a hybrid model that combines a Graph Attention Network (GCN) with a transformer to enhance skeleton-based human activity recognition (HAR). The model can selectively focus on important spatial-temporal features in skeleton action data by incorporating an attention mechanism, leading to more accurate recognition. This approach achieved a cross-subject accuracy of 89.70% and a cross-view accuracy of 95.70%, demonstrating its effectiveness in capturing local and global dependencies for improved performance in HAR tasks.

Wu et al. [14] proposed the Dual Stream Transformer GCN, a sophisticated model that integrates both temporal and spatial streams for enhanced human activity recognition (HAR). This dual-stream approach allows the model to simultaneously capture dynamic temporal changes and spatial relationships within the data, resulting in a more comprehensive understanding of human activities. The Dual Stream Transformer GCN achieved a cross-subject accuracy of 91.00% and a cross-view accuracy of 96.50%, underscoring its superior performance in addressing the complexities of HAR tasks through the effective combination of temporal and spatial features.

From the above literature review, it is observed that several advancements in human motion recognition provide a strong rationale for the proposed Efficient Net Conv LSTM (In short: EN Conv LSTM) method. Effective models, such as the Dual Stream Transformer GCN and various GCN-transformer hybrids, emphasise the importance of integrating spatial and temporal dynamics to achieve better accuracy, which EN Conv LSTM is designed to accomplish.

Additionally, approaches such as adaptive skeleton graphs and temporal edge aggregation demonstrate that flexibility in modelling is crucial for capturing human movements. The efficiency of the Efficient Net architecture, combined with Conv LSTM, enables more effective learning from motion data while reducing computational costs. Incorporating attention mechanisms can enhance EN Conv LSTM's ability to focus on critical features, thereby improving the recognition of subtle variations in motion patterns. Furthermore, the emphasis on geometric and structural features in existing models suggests that EN Conv LSTM can leverage its convolutional layers to capture these relationships, enhancing its effectiveness in recognizing actions based on joint and bone interactions. Lastly, the model's design aims to generalize across diverse scenarios, ensuring robust performance against variations in subjects and viewpoints. Overall, EN Conv LSTM is positioned to advance the field of human motion recognition by effectively integrating these insights, making it a promising model for achieving superior accuracy in human motion recognition systems.

III. METHODOLOGY

The proposed system begins by utilising this dataset to train the deep learning model, as shown in Figure 6. In the next step, the skeleton data is extracted from the raw video input, which contains joint positions and movements. Skeleton-based representations are used as input to the deep learning model [15]. This extraction simplifies the data, focusing on the skeletal movements of the human body to enhance the accuracy of motion detection. Then, the extracted skeleton data is processed by an Efficient Net model. Efficient Net is a convolutional neural network (CNN) known for its ability to extract spatial features from image sequences [16]. The key function of Efficient Net in this pipeline is to extract high-level spatial features from the skeleton data. Each block in Efficient Net refines the features while reducing the resolution of the input data. After the spatial features are extracted using EfficientNet, the next step involves temporal processing with an EfficientNet architecture [17]. Conv LSTM captures the temporal relationships between frames (i.e., human poses and movements change over time). This part of the model is crucial for detecting activities that have temporal dependencies, such as chest pain, headaches, vomiting, and other similar conditions. This hybridisation of EfficientNet and ConvLSTM is referred to as EN ConvLSTM. The extracted features from EfficientNet and ConvLSTM are used to train the model. The proposed methodology is presented in Figure 1. This figure represents an end-to-end system for human motion detection using a deep learning model, followed by real-time monitoring and notification.

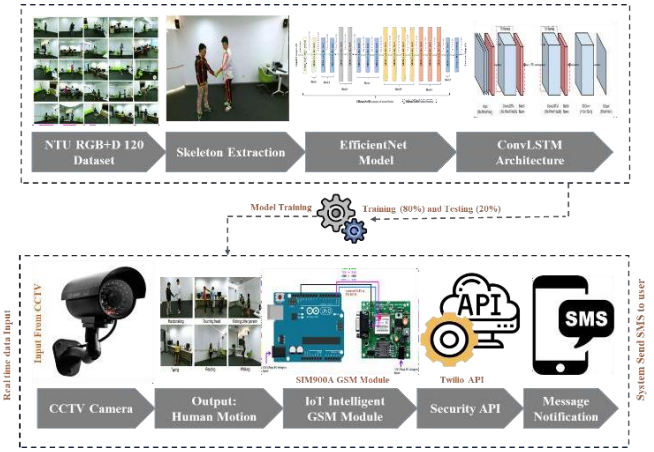
The dataset is split into training. (80%) and testing (20%) portions. The model is trained using 80% of the

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

data, and its performance is evaluated using the remaining 20%. During training, the model learns to differentiate between various human activities by adjusting its internal weights and parameters. During training, the model learns to recognise different human activities by minimising the classification loss, typically cross-entropy, using the Adam optimiser. The output of the trained EN Conv LSTM model represents the classified activity or motion. The classification loss function (L) is calculated using the following equation (1). selection criterion. The secondary criterion is the similarity of the video of the selected activities to those of the medical sector classes only. The

$$L = -\sum_{i=1}^{N} y_i * log\,(\hat{y}_i) \quad (1)$$

Where y_i is the proper level, $\hat{y}_i$is the predicted probability from The SoftMax output, and N is the number of classes. Once the model is trained and evaluated, it can be deployed for real-time human motion detection. A CCTV Camera is used to capture live video input, which is then processed by the system to detect human motions (such as chest pain, headache, vomiting, etc.). In the real-time monitoring system, the model outputs the detected human motion. This output is fed into an IoT Intelligent GSM Module (SIM900A GSM Module), which is responsible for communicating with a security API. This module allows the system to send notifications or alerts based on detected activities.



[Fig.1. The Workflow of the Proposed Methodology]

The IoT GSM module is connected to a Twilio API, which enables communication services, such as sending SMS messages. The Security API monitors detected activities and triggers notifications when model actions are detected. Based on the detection and the security API's output, the system sends an SMS to the user through the Twilio API. The notification alerts the user about the specific human motion detected, enabling real-time monitoring and action.

IV. EXPERIMENTAL SETUP

The quality and relevance of the dataset play a significant role in the overall performance of any type of machine learning or deep learning model, including the proposed network. That is why selecting an appropriate dataset based on the criteria determined by the experiment's goals is an essential step in Efficient Net Conv LSTM-based research. This subsection presents our process for choosing the most suitable dataset and methods for processing it.

third criterion is the feature richness of the available video dataset. We study and analyse six video datasets related to MHR, which are listed in Table I.

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

SN	Dataset Name	Total Samples Data	Total Classes
1	ActivityNet	21,313	200
2	Charades	66,493	157
3	HMDB51	6,766	51
4	NTU RGB+D 120	114,480	120
5	STAIR Actions	1,09,478	100
6	UCF101	13,320	101

A. Dataset Selection Criteria

This paper examines the feasibility of utilising existing HMR datasets, rather than creating a new one for the experiment. We primarily analyse three criteria when choosing the dataset. These criteria have been set to ensure the ease of implementing the proposed motion recognition system. That is why the availability of activities listed in Table I is a critical

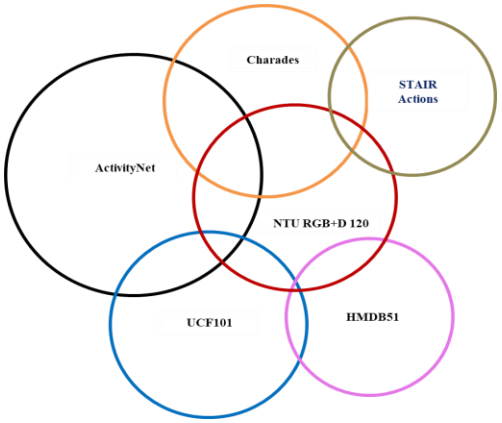


Fig. 2. Class overlapping among experimental datasets

We explore the skeleton datasets listed in Table II. However, it has been observed that there are overlapping classes among these datasets, as illustrated in Figure 2. The ActivityNet and Charades share the highest number of overlapping courses. That means these two datasets have a strong correlation with the "NTU RGB+D 120" dataset. The STAIR Actions dataset also shares some standard classes. The UCF101 and the HMDB51 datasets have the least standard courses. As a result, we conclude that these two datasets maintain a weak correlation with the "NTU RGB+D 120" dataset. Therefore, among these datasets, the "NTU RGB+D 120" dataset is considered in our experiment. Moreover, the current accuracy is insufficient for the "NTU RGB+D 120" dataset, as discussed in the literature review section. That is why we consider this dataset in our study.

B. NTU RGB+D 120 dataset

The "NTU RGB+D 120" dataset, developed by Nanyang Technological University, is a publicly available benchmark collection for human action recognition research [18]. "NTU RGB+D 120" Dataset is a large-scale dataset for human motion recognition that includes skeleton data. It consists of 120 classes with a total of

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

114,480 samples, featuring RGB videos (.mp4). At a resolution of 1920x1080 (Ratio: 16:9) pixels and a frame rate of 24 fps. Each sample also provides 3D coordinates of 25 body joints per frame. The sample of the dataset is presented in Figure 3. Our focus is on 12 specific medical classes in .mp4 data format. The classes include sneezing/coughing (950 samples), staggering (1050), chest pain (1100), and back pain. (1150), fan self (850), yawn (950), falling (1000), headache (900), neck pain (950), vomiting (1000), stretch oneself (1200), and blow nose (1100), totalling 13,200 samples. For both cross-subject and cross-view evaluations, the data is split into 80% training (10,560 samples) and 20% testing (2,640 samples), ensuring comprehensive coverage and robust assessment. Table II presents the selected 12 medical classes with their corresponding sample sizes.

Table II: The 12 Medical Classes with them

Action Number: Action/Class	No. of Skeletal Samples	Action Number: Action/Class	No. of Skeletal Samples
A41: sneeze/cough	950	A42: staggering	1050
A45: chest pain	1100	A46: back pain	1150
A49: fan self	850	A103: yawn	950
A43: falling down	1000	A44: headache	900
A47: neck pain	950	A48: vomiting	1000
A104: stretch oneself	1200	A105: blow your nose	1100
Total No. of samples	**6050**		**6150**
	13,200		



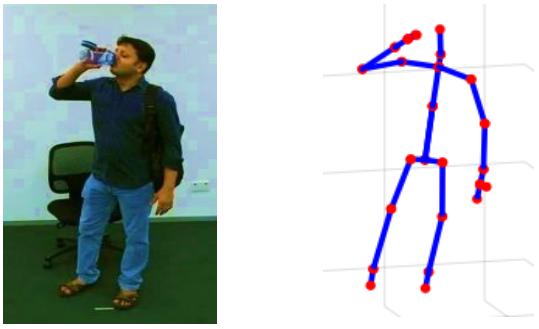
Fig 3. Sample Frames of the "NTU RGB+D 120" Dataset

☑ Cross-subject: The model's performance is evaluated by training it on data from a subset of subjects and testing it on data from different subjects. It assesses how well the model generalizes across various individuals.

☑ Cross-view: Cross-view refers to a scenario where a model is trained on data from one viewpoint or perspective (e.g., frontal view) and then tested or evaluated on data from a different viewpoint or perspective (e.g., side view).

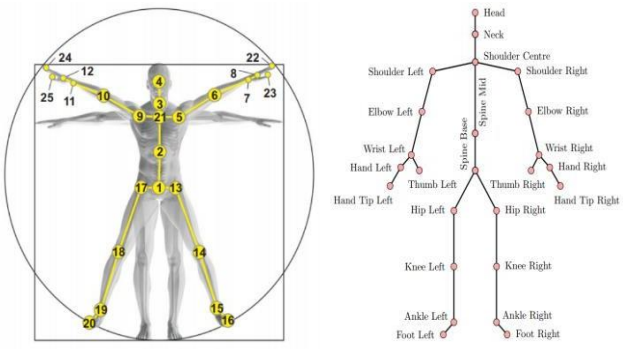
C. Data Preprocessing

The 25 joints of the human body, as detected by the Kinect version 2 camera, are shown in Figure 4. The numbering and names correspond to those used in the skeleton tracker. Joint information consists of the 3D locations of these major body joints for each detected and tracked human in the scene [19]. Pixels on RGB frames and depth maps are provided for each body joint. The configuration of these joints is illustrated in Figure 5. The RGB videos are recorded at a resolution of 1920 × 1080, while the infrared sequences are collected frame by frame at a resolution of 512 × 424. We employed skeleton joint estimation methods to obtain the body frames and implemented 3D joint tracking algorithms. Figure 5 shows the twenty-five (25) skeleton joints tracked at each instance. These joints include the right knee, right hip, left knee, left hip, left foot, left ankle, right foot, right ankle, head, mid spine, left wrist, right shoulder, left shoulder, right wrist, right elbow, and left elbow. All joint values contain X, Y, and Z coordinates in space. Each skeleton is extracted from the standard video at a frame rate of 24 frames per second (FPS), enabling frame-wise skeleton extraction. This 3D skeletal data contains 25 body joints at each frame. The skeleton data represents the human body's structure in three dimensions, consisting of 25 key body joints per frame. These joints form a skeletal representation that abstracts the human pose and movement, reducing the complexity of raw video data while preserving essential spatial and temporal features. The coordinates of each joint enable the system to analyse motion trajectories and interactions accurately, making the data suitable for training advanced models, such as EN Conv LSTM. By leveraging skeletal information, the system can focus on critical motion patterns, improving the efficiency and accuracy of human action recognition tasks.



(a) (b)

[Fig.4. (a) Human Activity (b) frame-wise skeleton extraction (24 frames per second)]



(a) (b)

[Fig. 5. Illustration of the configuration of 25 body joints in our dataset (a). The labels of these joints are (b): (1) base of spine. (2) middle of the spine, (3) neck, (4) head, (5) left shoulder, (6) left elbow, (7) left wrist, (8) left hand,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

(9) right shoulder, (10) right elbow, (11) right wrist, (12) right hand, (13) left hip, (14) left knee, (15) left ankle, (16) left foot, (17) right hip, (18) right knee, (19) right ankle, (20) right foot, (21) spine, (22) tip of the left hand, (23) left thumb, (24) tip of right hand, (25) right thumb]

Depends on this angle view. We can recognise the class levels. We extracted the human skeleton from the video data using the OpenPose Python library function [20]. The preprocessing algorithm for IoT-based real-time human motion recognition using EN Conv LSTM involves converting raw video data into .npy files to facilitate efficient model training and inference. The process begins with video

frame extraction, where frames are sampled at a specified rate (e.g., 10 frames per second) to capture the essential motion patterns without redundant data. Each frame is resized to a consistent resolution to ensure compatibility with the EN Conv LSTM model architecture [21]. Additional preprocessing steps, such as grayscale conversion and pixel normalisation, are applied to enhance feature representation and accelerate model convergence. Once the frames are processed, they are organised into sequential arrays that represent the temporal dynamics of human motion. These arrays are saved in the .npy format, a lightweight and efficient data storage format in NumPy, ensuring the structured preservation of frame sequences for rapid access during training and real-time predictions.

V. PROPOSED ARCHITECTURE

The proposed model, EN Conv LSTM, aims to enhance real-time Human Motion Recognition (HMR) by leveraging the strengths of EfficientNet and ConvLSTM networks. This hybrid approach addresses the limitations of existing methods, such as high computational complexity, low accuracy, and limited adaptability to diverse healthcare scenarios.

A. EN Conv LSTM Proposed Model Architecture

Figure 6 presents a hybrid architecture for human motion recognition, combining EfficientNet for spatial feature extraction and ConvLSTM for temporal feature modelling. At the top, the EfficientNet architecture is shown, which processes input frames (224x224 resolution) through a series of MBConv blocks. Once the skeleton data is extracted, it is passed through the EfficientNet model for spatial feature extraction. EfficientNet uses a series of Mobile Inverted Bottleneck Convolutions (MBConv) to downsample and extract meaningful features from the input. The structure consists of multiple blocks of convolutions, where the operations can be represented as the equation (2).

$$y = f(x; W) = MBConv(x; W) \quad \dots \quad (3)$$

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. 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, resulting in high-level representations for the next stage.

Forget gate:

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f) \quad \dots \quad (4)$$

Input gate:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i) \quad \dots \quad (5)$$

Cell state update:

$$C = f \odot C + i \odot \tanh(W * X) + W * H + b \quad \dots \quad (6)$$

an AMD Ryzen 9 5900X 12-Core Processor, running on a 64-t t t−1 t xc t hc t−1 c

bit system with Python 3.9.13 and CUDA 11.0. It includes an *Input gate:*

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o) \quad \dots \quad (7)$$

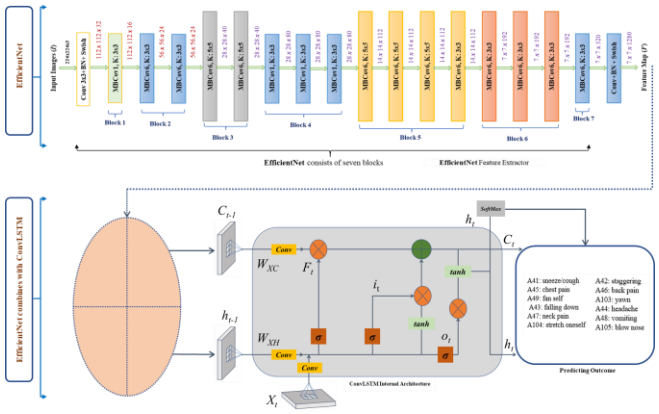
Hidden state: $H_t = o_t \odot \tanh(C_t)$ ... *(8)*

Here's a breakdown of the symbols and their meanings used in the equations:

- ☑ f_t: Forget gate activation at time step *t*
- ☑ i_t: Input gate activation at time step *t*
- ☑ o_t: Output gate activation at time step *t*
- ☑ C_t: Cell state at time step *t*
- ☑ H_t: Hidden state (or output) at time step *t*
- ☑ σ: Sigmoid activation function
- ☑ *tanh*: Hyperbolic tangent activation function
- ☑ $W_x f$, $W_x i$, $W_x c$, $W_x o$: Weight matrices for the input X_t in each gate (forget gate, input gate, cell state update, output gate)
- ☑ $W_h f$, $W_h i$, $W_h c$, $W_h o$: Weight matrices for the hidden state H_{t-1} (from the previous time step) in each gate
- ☑ $b_x f$, $b_x i$, $b_x c$, $b_x o$: Bias terms for each gate (forget gate, input gate, cell state update, output gate)
- ☑ ⊙: Hadamard product (element-wise multiplication)
- ☑ X_t: Input at the current time step *t*
- ☑ H_{t-1}: Hidden state (output) from the previous time step *t-1*

These symbols represent how information flows and is manipulated within the ConvLSTM (Convolutional Long Short-Term Memory) architecture, where information is selectively remembered or forgotten over time through these gates. This architecture outputs temporal features, which are classified into predefined activity classes using a SoftMax layer, as shown in Figure 5. The system can recognise various actions, such as sneezing, yawning, or back pain (A41–A105 classes). This approach efficiently captures both spatial and temporal dependencies, making it highly effective for human motion recognition tasks.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



[Fig.6: The Internal Structure of the EN Conv LSTM Block]

VI. RESULT AND DISCUSSION

A. Software and Hardware Requirements

The study requires several key software and hardware components. The software incorporates deep learning frameworks, including TensorFlow, PyTorch, scikit-learn, and Keras, along with data processing libraries such as NumPy, Pandas, and OpenCV, all of which utilise Python. Development is done in Jupyter Notebook and PyCharm. The hardware setup features 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, ensuring a compact and versatile integration of hardware and software for deep learning and data processing.

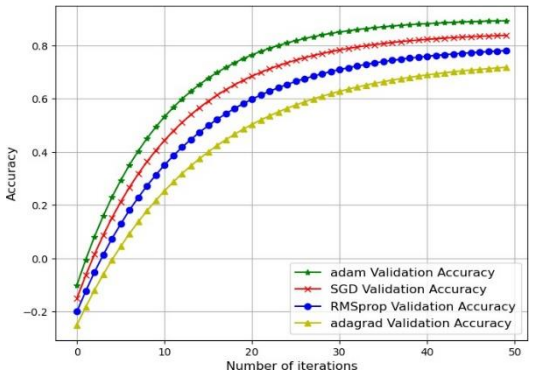
B. Model Parameters List

The ENConvLSTM model has a total of 8,138,104 parameters, with 8,095,640 being trainable and 42,464 non-trainable. The model is configured to use a batch size of 16 and will be trained for 100 epochs. The loss function employed is sparse categorical cross-entropy, while the optimizer used is Adam. The activation function is set to SoftMax, and the learning rate is established at 0.001. The ENConvLSTM Model Parameters are listed in Table III.

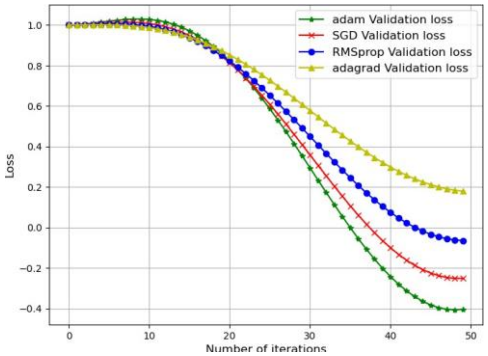
Table-III: ENConvLSTM Model Parameters

SN	Model Parameters	Value Count and Function
1	Total Parameter	8138104
2	Trainable Parameter	8095640
3	Non-trainable Parameter	42464
4	Batch Size	16
5	Epoch	100
6	Loss function	Sparse categorical crossentropy
7	Optimizer	Adam
8	Activation Function	SoftMax
9	Learning Rate	0.001

Figure 7 illustrates the superior performance of the Adam optimiser compared to SGD, RMSprop, and Adagrad across 50 epochs of training the proposed model. Adam consistently achieves the highest validation accuracy, approaching 0.95, while also maintaining the lowest validation loss throughout the training process. This is attributed to Adam's adaptive learning rate mechanism, which tailors the learning rate for each parameter based on the moving averages of the gradients, resulting in faster convergence and improved generalisation. The stability and efficiency of Adam, as demonstrated by the smooth and less noisy accuracy and loss curves, further underscore its effectiveness in optimizing the model. In contrast, SGD, RMSprop, and Adagrad show slower convergence and higher validation losses, indicating less effective optimization.



(a)



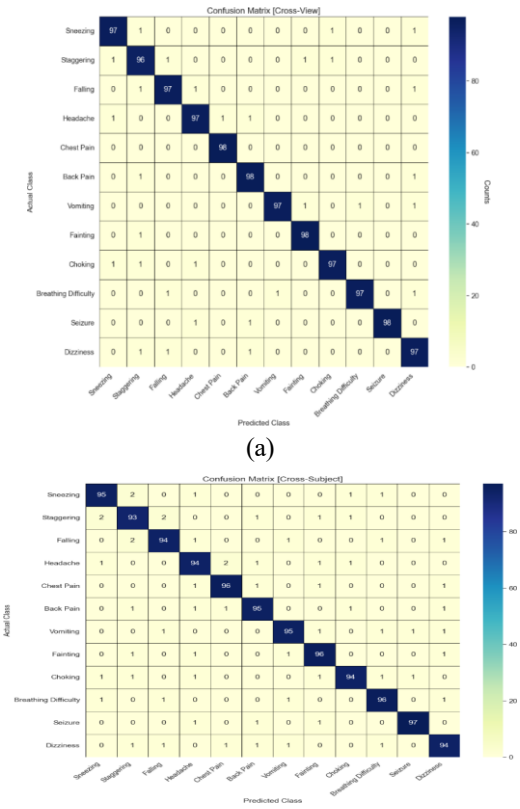
(b)

[Fig.7. Model Optimizers Comparison (a) Validation Accuracy and (b) Validation loss]

Overall, the results strongly suggest that Adam is the most suitable optimiser for this model, offering a balanced approach to both speed and accuracy, making it an ideal choice for complex deep learning tasks in healthcare scenarios. We analysed the validation loss curve over the number of iterations, as illustrated in Figure 7. It shows that Adam performs better than the AdaGrad and RMSProp algorithms.

Figures 8-10 present the model's performance across various evaluation metrics. Figure 8 illustrates the confusion matrix for Cross-View and the model's loss for Cross-Subject. In Figure 9, we compare the training and validation accuracy (a) and loss (b) for the Cross-Subject evaluation. The Figure indicates that the model is learning effectively, with increasing accuracy and decreasing loss for both the training and validation sets. There are no signs of overfitting, as the performance on the validation set closely aligns with that on the training set. Similarly, Figure 10 provides the training and validation accuracy (a) and loss (b) for the Cross-View scenario. These figures comprehensively demonstrate the model's performance across different perspectives and evaluation setups.

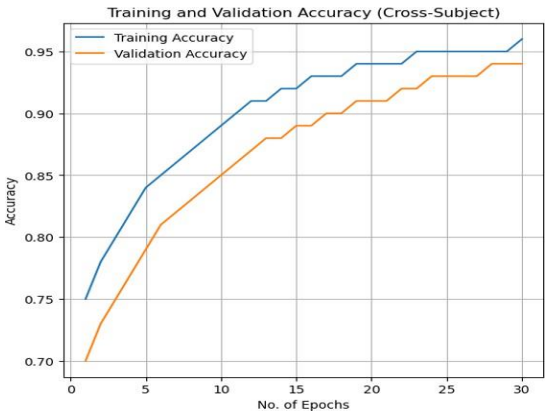
Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



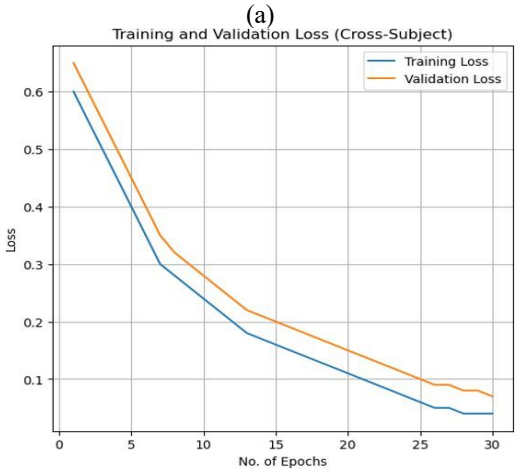
(a)

(b)

Fig. 8. Confusion Matrix (Cross-View) and (b) Model Loss (Cross-Subject)



(a)



(b)

Fig. 9. Model Accuracy. Training and Validation Accuracy (a) and (b). Training and Validation loss for Cross-Subject

C. Result Evaluation Methods

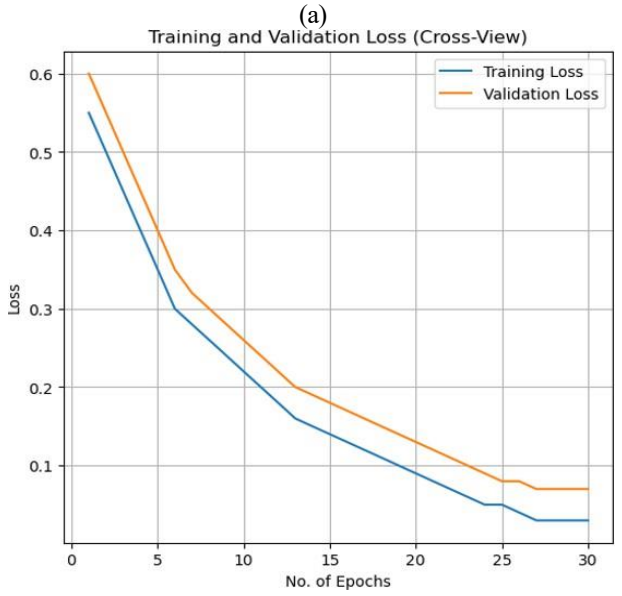
The overall accuracy of this dataset is listed in Table IV. And Its Result in Comparison with Other Used Models. This comparison highlights the strength of your ENConvLSTM model in Human Motion Recognition (HMR) tasks, particularly in improving overall accuracy across both challenging metrics.

☑ **Cross-Subject Evaluation**: This involves splitting the dataset so 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%.

☑ **Cross-View Evaluation**: This involves splitting the dataset based on different camera views to ensure the model performs well across various perspectives. The model achieved a cross-view accuracy of 96.45%.



(a)



(b)

Fig. 10. Model Accuracy, Training and Validation Accuracy, and (b) Training and Validation Loss for Cross-View

Retrieval Number: 100.1/ijitee.D105514040325
DOI: 10.35940/ijitee.D1055.12030223
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Table-IV: Result in Comparison with Other Used Models (For Overall Accuracy)

Previous Used Methods	Cross-Subject (%)	Cross-View (%)
Dynamic Skeleton [22]	60.2	65.2
P-LSTM [23]	62.9	70.3
STA-LSTM [24]	73.4	81.2
TCN [25]	74.3	83.1
VA-LSTM [26]	79.2	87.7
Deep STGCK [27]	74.9	86.3
ST-GCN [28]	81.5	88.3
DPRL [29]	83.5	89.8
SR-TSL [30]	84.8	92.4
STGR-GCN [31]	86.9	92.3
AS-GCN [32]	86.8	94.2
2S-AGCN [33]	88.5	95.1
2sAGC-LSTM [34]	89.2	95.0
NAS-GCN [35]	89.4	95.7
ST-GCN [36]	90.1	95.9
Ours (ENConvLSTM)	**94.85**	**96.45**

D. Result Performance Metrics

The proposed ENConvLSTM model is evaluated using the "NTU RGB+D 120" dataset, focusing on 12 selected medical classes. The results demonstrate the model's effectiveness in real-time Human Motion Recognition (HMR) for healthcare applications. Key performance metrics, including accuracy, precision, recall, and F1-score, are calculated for both cross-subject and cross-view evaluations, as listed in Table V. The visual presentation of this table is shown in Figure 11. The model performs exceptionally well across all evaluation metrics, particularly for cross-view, which typically involves more variability (e.g., different camera angles). These results demonstrate the robustness and generalizability of the model on the NTU RGB+D 120 dataset.

The results are also compared with other state-of-the-art models to highlight the improvements achieved by the ENConvLSTM model. The Result Performance Metrics are written in equations number from 9 to 12.

Table-V: Other Performance Evaluation Metrics on the "NTU RGB+D 120" Dataset

SN	Classification 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%

E. Accuracy and Loss Summary

Below is a summary of Table VI, which presents the accuracy and loss for both cross-view and cross-subject evaluations, along with the corresponding hypothetical accuracy and loss curves.

Table-VI: Accuracy and Loss Summary

	Cross-Subject	Cross-View
Accuracy Summary	• Initial Accuracy: 75.01% • Final Accuracy: 94.85%	• Initial Accuracy: 78.34% • Final Accuracy: 96.45%
Loss Summary	• Initial Loss: 0.60 • Final Loss: 0.04	• Initial Loss: 0.55 • Final Loss: 0.03

In the cross-subject evaluation, accuracy exhibits a robust upward trend, increasing from an initial 75.01% to a final 96.85%, with some minor fluctuations and eventual stabilisation. 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.

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

indicating effective learning and convergence. Both evaluations demonstrate an effective model performance enhancement over time, with accuracy improving and loss decreasing.

$$Precision_{PerClass} = \quad (10)$$

decreasing, culminating in stable performance metrics.

$$Recall$$

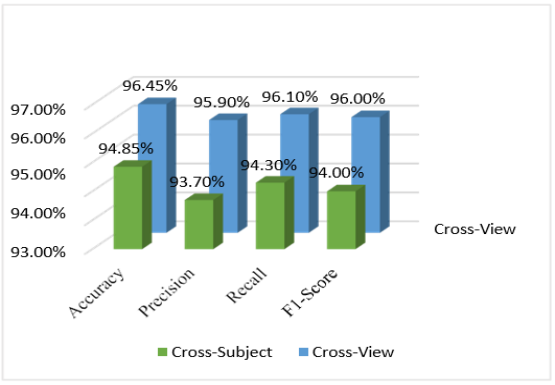
$$PerClass = \frac{TP}{TP+FP} \quad (11)$$

F. Result Comparison

The performance of the ENConvLSTM model is compared.

$$F1-score_{PerClass} = \frac{2*TP}{[2*TP+FP+FN]} \quad (12)$$

utilising several state-of-the-art models, including traditional methods and recent deep learning approaches. The comparative analysis demonstrates the significant improvements achieved by the proposed model. The performance of the ENConvLSTM model is compared with several state-of-the-art models, including traditional methods and recent deep-learning approaches. The comparative analysis, as shown in Table VII, demonstrates that the proposed model achieves significant improvements. Each Performance Evaluation Method (Cross-Subject and Cross-View) is graphically presented in Figure 12. 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.

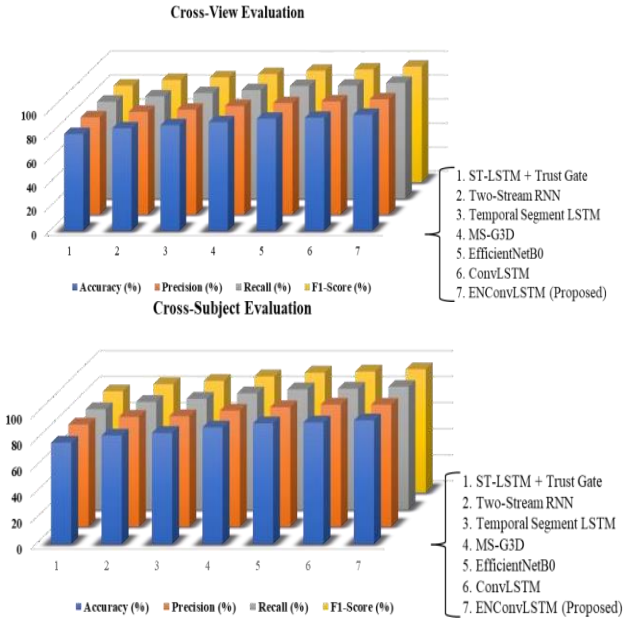


[Fig. 11. Model Performance Evaluation on "NTU RGB+D 120" Dataset]

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Table VII: "NTU RGB+D 120" Dataset Accuracy Summary List

Used Model with citation	Evaluation Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ST-LSTM + Trust Gate [37]	Cross-Subject	77.71	78.54	77.14	77.81
Two-Stream RNN [38]		83.22	84.43	82.76	83.32
Temporal Segment LSTM [39]		85.23	84.91	85.18	85.77
MS-G3D [40]		89.44	89.15	89.33	89.23
EfficientNetB0 [41]		92.55	91.86	92.45	92.15
ConvLSTM [42]		93.23	93.78	92.96	92.95
ENConvLSTM (Proposed)		**94.85**	**93.72**	**94.31**	**94.76**
ST-LSTM + Trust Gate [37]	Cross-View	80.47	80.91	80.54	80.48
Two-Stream RNN [38]		85.24	85.54	84.95	85.25
Temporal Segment LSTM [39]		87.82	87.26	87.63	87.42
MS-G3D [40]		90.25	90.31	90.15	90.05
EfficientNetB0 [41]		93.17	92.73	93.54	92.85
ConvLSTM [42]		94.02	94.16	93.84	93.95
ENConvLSTM (Proposed)		**96.45**	**95.91**	**96.17**	**96.32**



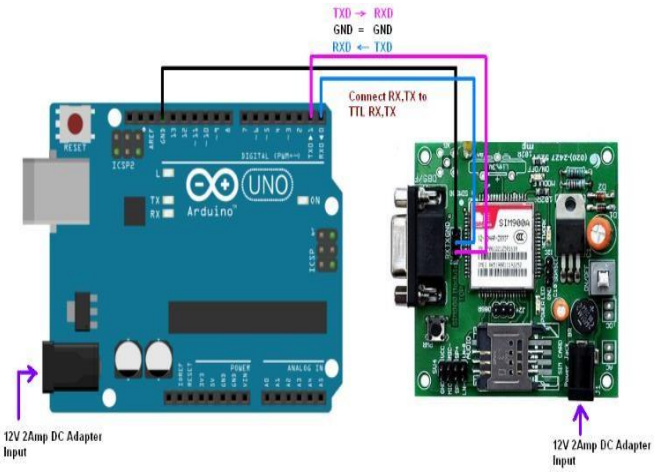
[Fig.12: Performance Comparison Among Different Tested Models on the "NTU RGB+D 120" Dataset]

VII. REAL-TIME TESTING

A. Real-Time Implementation

To implement the real-time scenario, Figure 13 illustrates the interconnection between an Arduino Uno and the SIM900 GSM module. A 12V 2A DC adapter powers the Arduino. The TX (transmit) and RX (receive) pins are crucial for communication. The TX pin of the Arduino is connected to the RXD pin of the TTL module for data transmission, while the RX pin of the Arduino is connected 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 such as sending SMS messages or making voice calls.

Figure 14 illustrates the wiring connections between a Raspberry Pi and an ESP8266 Wi-Fi module using a breadboard. 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.

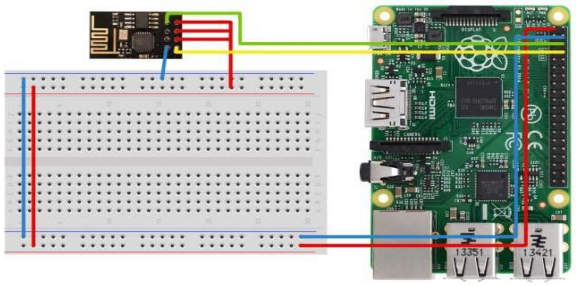


[Fig. 13: Schematic Diagram Illustrating the Connections between an Arduino UNO and a SIM900 GSM TTL Module, Highlighting the Power Supply, TX and RX pin Connections, and the Common Ground to Facilitate Communication for Tasks Such as SMS sending and Voice Calling]

The yellow wire connects the GPIO pin on the Raspberry Pi to the ESP8266's TX pin for data transmission, and the green wire connects another GPIO pin to the ESP8266's RX pin for data reception. This setup enables the Raspberry Pi to communicate with the ESP8266 for wireless connectivity in various projects, such as IoT applications, using the Arduino UNO REV3, a compact and versatile microcontroller board (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. Figure 15 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 expected 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.

Retrieval Number: 100.1/ijitee.D105514040325
DOI: 10.35940/ijitee.D1055.12030223
Journal Website: www.ijitee.org

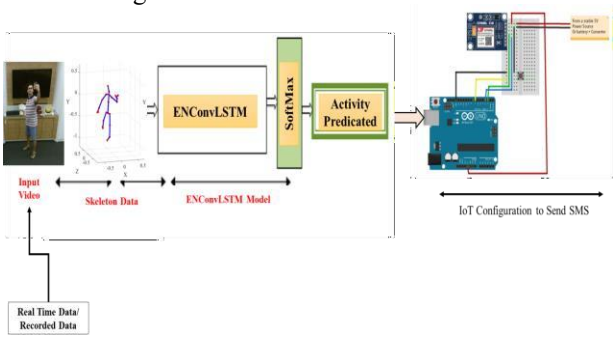
Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



[Fig. 14. Schematic Diagram Showing the Wiring Connections Between A Raspberry Pi and an ESP8266 Wi-Fi Module Using A Breadboard, Illustrating Power (VCC And GND) and Data Transmission (TX And RX) Connections for Enabling Wireless Communication in IoT Projects]

This setup enables the effective monitoring and recognition of human activities through skeletal motion analysis. The proposed human motion recognition system's

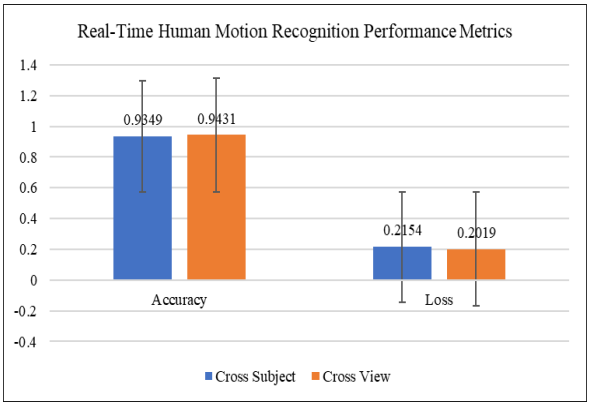
Performance in the laboratory experiment is satisfactory. However, there are always some differences between laboratory and real-time scenarios. Table VIII illustrates the results in a real-time scenario, and their visual representation is shown in Figure 16.



[Fig. 15: Diagram Illustrating a Human Motion Recognition System Using an Enconvlstm Model, Showcasing the Process of Extracting Skeleton Data from the Input Video, Predicting Activity Via a Softmax Layer, and Communicating the Results to an Arduino for Real-Time or Recorded Data Applications]

Table VIII: Real-Time Human Motion Recognition Performance Metrics (Includes Accuracy and Loss Values)

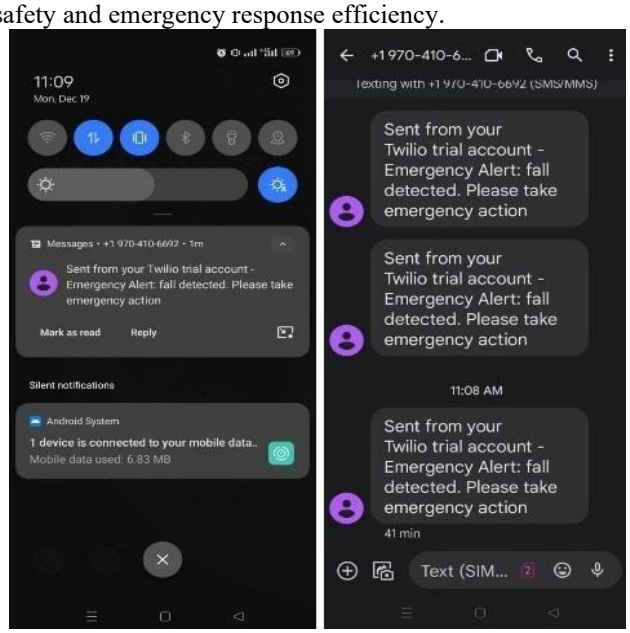
Evaluation	Accuracy	Loss
Cross Subject	93.49%	0.2154
Cross View	94.31%	0.2019



[Fig. 16: Real-Time Human Motion Recognition Performance Metrics (Includes Accuracy and Loss Values)]

B. 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's mobile device. Figure 17 presents snapshots of the warning messages generated by the system when it detects a fall event. This real-time communication is facilitated through the integration of Twilio, 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.



[Fig.17: Snapshots of Warning Messages]

VIII. LIMITATION AND FUTURE RESEARCH SCOPE

The proposed system's limitations are summarized in the following section. This limitation suggests that further research is needed to enhance the system's robustness.

A. Subject and Camera Position

One limitation of the proposed methodology is subject sensitivity. It has been observed that if the subject is too close to the camera (approximately 15 cm), it can adversely affect the system's performance. If the subject is too far from the camera (about 25m), the system can't detect human activity. A subject position too close to the camera (approximately 15 cm) or too far (around 25 meters) might introduce variability that can affect the model's performance. Typically, maintaining a consistent and appropriate distance is crucial for reliable and accurate analysis. For optimal results, subjects should be within a range where the camera can capture relevant details without distortion or loss of information.

B. Night-Vision Experiment

The suggested methodology is tested in the presence of light in the room or during daylight. Trying out night vision mode is an additional issue that is not covered in

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

this study. Subsequent studies, however, are being conducted to investigate the performance in night vision mode.

C. Class Activity Limitation

In this paper, we consider only 12 medical classes. There are also many classes related to the medical sector that we do not include because of data availability. A customised dataset is designed for our proposed algorithm. It paves the way for conducting more research, including creating new datasets and processing them, discovering limitations and improving them for new classes, and handling computational complexities for additional
courses. The limitations of the proposed algorithm pave the way for further development of this system. Instead of considering these as limitations, the researchers of This study consider them as opportunities to continue the research work and integrate more innovative features with it.

D. Environmental Conditions

Varying environmental conditions, including lighting, background clutter, and weather conditions, can significantly influence the performance of the proposed system. For instance, sudden changes in illumination or highly cluttered backgrounds may lead to misclassification or an inaccurate detection of activities. Future research should focus on developing methods to make the system more resilient to environmental changes and capable of operating reliably under diverse conditions.

E. Real-time Processing and Latency

While the proposed system aims for real-time HMR, it may encounter challenges related to processing speed and latency, particularly when handling high-resolution videos or complex activities. Ensuring real-time performance without compromising accuracy requires optimising the algorithm and possibly leveraging hardware accelerators, such as GPUs or specialised AI processors. Future research could focus on improving the computational efficiency and reducing latency to make the system more suitable for real-time applications.

F. User Adaptability and Customization

The current system may not account for individual differences in performing activities, which can vary significantly among users due to factors such as age, physical condition, or personal style. Incorporating user-specific adaptations and customization features can improve the system's accuracy and user satisfaction. Future research could explore adaptive algorithms that learn and adjust to individual user behaviors over time.

G. Computational Load

Combining multiple deep learning architectures, such as EfficientNet, LSTM, and ConvLSTM, into a single model named ENConvLSTM for real-time Human Motion Recognition introduces several factors that can impact model complexity and computational load. This integration increases the number of parameters and the computational burden, necessitating the use of high-performance GPUs to handle the load efficiently. Ensuring that the hardware setup, including memory and processing power, can accommodate these demands is crucial for maintaining real-time performance.

IX. CONCLUSION AND FUTURE DIRECTION

This paper presents an innovative IoT-based solution for real-time Human Motion Recognition (HMR) in healthcare, utilising the ENConvLSTM deep learning model, which combines EfficientNet for spatial feature extraction and ConvLSTM for temporal data processing. The use of skeletal data enables the precise detection of subtle motion patterns, which are essential for early diagnosis and intervention in healthcare scenarios. A key contribution of the study is the proposed hybrid deep learning model, which improves the accuracy rate to 94.85% for cross-subject and 96.45% for cross-view, effectively addressing the challenges posed by the “NTU RGB+D 120” public dataset. Another key contribution is the integration of the model with a Raspberry Pi microcontroller and GSM module, enabling real-time notifications for healthcare providers and patients, and facilitating proactive monitoring and immediate responses without the need for third-party mobile applications. The model's ability to recognise 12 medically relevant classes demonstrates its adaptability to various healthcare scenarios, making it a practical and scalable solution. This research significantly advances HMR by providing a robust, efficient model that integrates seamlessly with IoT frameworks, with the potential to enhance patient care, reduce healthcare costs, and enable more responsive healthcare systems. The ENConvLSTM model’s success in addressing both spatial and temporal aspects of data sets the foundation for future work in leveraging deep learning and IoT for advanced healthcare applications.

In future, there are several promising directions to enhance the performance and applicability of the proposed IoT-based real-time Human Motion Recognition (HMR) system. One potential avenue is to further improve the accuracy and generalisation of the ENConvLSTM model by incorporating advanced data augmentation techniques or experimenting with newer deep learning architectures, such as Vision Transformers (ViTs), for enhanced spatial-temporal feature extraction. Additionally, expanding the model to handle more complex medical classes and different types of sensor data, such as integrating RGB video and depth data, could improve its adaptability to diverse healthcare scenarios. Another future direction involves optimizing the system for deployment in resource-constrained environments by exploring lightweight models or model compression techniques, such as pruning and quantization, to reduce computational overhead while maintaining high accuracy.

This would make the system more feasible for real-time applications on embedded devices, such as wearable healthcare monitors. Furthermore, integrating the model with cloud computing services could enable large-scale remote monitoring and storage of patient data, providing a scalable solution for broader healthcare applications. Investigating privacy and security aspects, particularly in the context of sensitive patient data, is also essential

© Copyright: All rights reserved.

for deploying such systems in real-world healthcare environments. Lastly, future research could focus on refining the real-time alert system by incorporating AI-driven anomaly detection methods to predict critical health issues before they fully manifest. This proactive approach could further enhance patient care by providing early warnings and reducing the incidence of emergencies. Collaborations with medical professionals and healthcare providers will be crucial for validating and refining the system for practical, real-world deployment in clinical settings.

ACKNOWLEDGMENT

I extend my sincere thanks to the Information and Communication Technology Division of the Ministry of Posts, Telecommunication, and Information Technology of Bangladesh, People's Republic of Bangladesh, for their invaluable support and funding of my ICT fellowship program in the MPhil Degree. Additionally, I would like to express my gratitude to my supervisor and co-authors for their guidance and contributions to this research.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** I extend my sincere thanks to the ICT Division of the Ministry of Posts, Telecommunications, and Information Technology of the People's Republic of Bangladesh for their invaluable support and funding of my ICT fellowship program within the MPhil program. The research Grant Number is 56.00.0000.052.33.005.21-7 (Tracking No: 22FS15306) and is supported by the University of Rajshahi.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** Md does the conceptualization of the research. Ekramul Hamid. Subrata Kumer Paul develops the methodology. Subrata Kumer Paul also handles data collection and preprocessing. Rakhi Rani Paul, Subrata Kumer Paul and Md. Ekramul Hamid used the necessary software for the project. Subrata Kumer Paul and Rakhi Rani Paul wrote the original draft and the final copy. Throughout the project, Md. Ekramul Hamid provides overall supervision, ensuring coherence and adherence to project goals, which stipulates the involvement of each author. The level of detail differs; some subjects yield articles that consist of isolated efforts that can be easily detailed, while other areas function as group efforts at all stages.

REFERENCES

1. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Advances in Neural Information Processing Systems (NIPS), 802-810. https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf
2. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), 6105-6114. https://doi.org/10.48550/arXiv.1905.11946
3. World Health Organization. (2019). WHO Global Report on Falls Prevention in Older Age. https://apps.who.int/iris/handle/10665/43811.
4. N. Heidari and A. Iosifidis, "Progressive spatio-temporal graph convolutional network for skeleton-based human action recognition," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3220–3224. https://doi.org/10.48550/arXiv.2011.05668
5. P. Geng, X. Lu, W. Li, and L. Lyu, "Hierarchical aggregated graph neural network for skeleton-based action recognition," IEEE Transactions on Multimedia. https://research-repository.uwa.edu.au/en/publications/hierarchical-aggregated-graph-neural-network-for-skeleton-based-a
6. Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," IEEE Transactions on Multimedia, vol. 25, pp. 1819–1831, 2023. https://dl.acm.org/doi/abs/10.1109/TMM.2022.3168137
7. L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Adaptive spectral graph convolutional networks for skeleton-based action recognition," 2018. http://dx.doi.org/10.48550/arXiv.1805.07694
8. C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021, ser. Lecture Notes in Computer Science, e., Del Bimbo, A., Ed. Springer, Cham, 2021, vol. 12663. https://doi.org/10.1007/978-3-030-68796-0_50
9. Y. Song, J. Yuan et al., "Temporal-aware adaptive skeleton graph network for skeleton-based action recognition," in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 3248–3252. http://dx.doi.org/10.1145/3484274.3484288
10. Y. Duan, H. Peng et al., "Hybrid model for spatio-temporal skeleton action recognition," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. pp. 7658–7667.
11. D. Zhao et al., "Learning skeleton aware geometry features for action recognition," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1504–1514.
12. S. Yang, C. Zhao et al., "Temporal edge aggregation for skeleton-based action recognition," in 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 3370–3374. https://doi.org/10.48550/arXiv.2205.15936
13. X. Feng et al., "Graph attention network and transformer for skeleton action recognition," in 2023 IEEE International Conference on Multimedia and Expo (ICME), 2023, pp. 890– 895.
14. Z. Wu et al., "Dual stream transformer GCN for skeleton-based action recognition," in 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.
15. Chuan-Bi Lin, Ziqian Dong, Wei-Kai Kuan, Yung-Fa Huang, "A Framework for Fall Detection Based on OpenPose Skeleton and LSTM/GRU Models," Applied Sciences Vol. 11, No. 329, 2021. https://doi.org/10.3390/app11010329
16. Prathima P, Dr. Ravindra S, "Detecting Real-Time Fall of Elderly People Using Machine Learning" International Journal for Research in Applied Science & Engg Technology Vol. 9, Dec 2021, 1913-1918. http://dx.doi.org/10.1109/ICCCNT49239.2020.9225494
17. Gangireddy Prabhakar Reddy, M. Kalaiselvi Geetha, "Video Based Fall Detection Using Deep Convolutional Neural Network." European Journal of Molecular & Clinical Medicine, 18 December 2020. https://rose1.ntu.edu.sg/dataset/actionRecognition/
18. World Health Organization. (2019). WHO Global Report on Falls Prevention in Older Age. https://apps.who.int/iris/handle/10665/43811
19. Chuan-Bi Lin, Ziqian Dong, Wei-Kai Kuan, Yung-Fa Huang, "A Framework for Fall Detection Based on OpenPose Skeleton and LSTM/GRU Models," Applied Sciences Vol. 11, No. 329, 2021.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

https://doi.org/10.3390/app11010329
20. Prathima P, Dr. Ravindra S, "Detecting Real-Time Fall of Elderly People Using Machine Learning" International Journal for Research in Applied Science & Engg Technology Vol. 9, Dec 2021, 1913-1918. http://dx.doi.org/10.1109/ICCCNT49239.2020.9225494
21. Gangireddy Prabhakar Reddy, M. Kalaiselvi Geetha, "Video-Based Fall Detection Using Deep Convolutional Neural Network." European Journal of Molecular & Clinical Medicine, 18 December 2020. https://doi.org/10.1109/ISCAS.2018.8351648
22. Jisu Kim and Deokwoo Lee: Activity Recognition with Combination of Deeply Learned Visual Attention and Pose Estimation. Appl. Sci. 2021, 11, 4153. https://doi.org/10.3390/app11094153
23. Han Chen, Yifan Jiang, And Hanseok Ko: Pose-Guided Graph Convolutional Networks for Skeleton-Based Action Recognition. IEEE Access (Volume: 10), 2022, http://doi.org/10.1109/ACCESS.2022.3214812
24. Wei Peng, Jingang Shi, Tuomas Varanka, Guoying Zhao: Rethinking the ST-GCNs for 3D skeleton-based human action recognition. Neurocomputing Volume 454, 24 September 2021, Pages 45-53, https://doi.org/10.1016/j.neucom.2021.05.004
25. J.-W. Hsieh, S.-H. Yu, Y.-S. Chen, and W.-F. Hu, "Automatic traffic surveillance system for vehicle tracking and classification," IEEE Transactions on Intelligent Transportation Systems, vol. 7, pp. 175–187, 2006. https://doi.org/10.1109/TITS.2006.874722
26. Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossings for driving assistance systems," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 3, pp. 729–739, 2012. https://doi.org/10.1109/ICITAMEE50454.2020.9398493
27. M. Sreeraj, "Multi-posture human detection based on hybrid HOG-BO feature," in Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on, pp. 37–40, Kochi, India, 2015. https://doi.org/10.1109/ICACC.2015.99
28. H. Liu, Z. Ju, X. Ji, C. S. Chan, and M. Khoury, "Study of human action recognition based on Human Motion Sensing and Recognition, pp. 233–250, Springer, 2017. http://dx.doi.org/10.1007/978-3-662-53692-6_11
29. R. M. Haralick, "Statistical and structural approaches to texture," Proceedings of the IEEE, vol. 67, pp. 786–804, 1979. https://doi.org/10.1109/PROC.1979.11328
30. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in Matters of Intelligence, pp. 115–141, Springer, 1987. https://pubmed.ncbi.nlm.nih.gov/3836989/
31. Z. Gao, H. Zhang, A. A. Liu, G. Xu, and Y. Xue, "Human action recognition on depth dataset," Neural Computing and Applications, vol. 27, no. 7, pp. 2047–2054, 2016. https://doi.org/10.1007/s00521-015-2002-0
32. S. Aly, "An effective human action recognition system based on Zernike moment features," in 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), pp. 52–57, Aswan, Egypt, 2019. https://doi.org/10.1109/ITCE.2019.8646504
33. D. K. Vishwakarma, "A two-fold transformation model for human action recognition using decisive pose," Cognitive Systems Research, vol. 61, pp. 1–13, 2020. https://doi.org/10.1016/j.cogsys.2019.12.004
34. Z. Chen, C. Jiang, S. Xiang, J. Ding, W. Min, and X. Li, "Smartphone sensor-based human activity recognition using feature fusion and maximum a posteriori," IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 7, pp. 3992– 4001, 2020. https://doi.org/10.1109/TIM.2019.2945467
35. S. Salahuddin, I. Ahmed, M. Rashid, and N. Minallah, "Automatic recognition of human actions," in 2019 13th International Conference on Open Source Systems and Technologies (ICOSST), pp. 1–6, Lahore, Pakistan, 2019. https://doi.org/10.1109/ICOSST48232.2019.9043921
36. W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), vol. 1, pp. 719–722, IEEE, 2004. https://doi.org/10.1109/ICME.2004.1394293
37. Lv and R. Nevatia, "Single view human action recognition using key pose matching and Viterbi path searching," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp. 1–8, Minneapolis, MN, USA, 2017. https://doi.org/10.1109/CVPR.2007.383131
38. D. Tran and A. Sorokin, "Human activity recognition with metric learning," in European conference on computer vision, pp. 548–561, Berlin, Heidelberg, 2008. https://doi.org/10.1007/978-3-540-88682-2_42
39. Q. Wang, D. Gong, M. Li, C. Zhao, and Y. Lei, "Sparse feature auto-combination deep network for video action recognition," in 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 712–716, Guilin, China, 2017. https://doi.org/10.1109/FSKD.2017.8393360
40. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 886–893, San Diego, CA, USA, 2015. https://doi.org/10.1109/CVPR.2005.177
41. Aggarwal, J. K., & Ryoo, M.S. (2011). Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3), 1-43. https://doi.org/10.1145/1922649.1922653
42. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2625-2634. https://doi.org/10.1109/TPAMI.2016.2599174

AUTHORS PROFILE



Subrata Kumer Paul completed his B.Sc. and M.Sc. Engineering from Rajshahi University, in Computer Science and Engineering in 2016 and 2017, respectively. Currently, he is working as an Assistant Professor at the Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad Cantonment, Natore-6431, Bangladesh. Also, he has been pursuing. The MPhil Program at Rajshahi University. His research fields are Deep Learning, Data Mining, and Speech Signal Processing. He has published more than 25 international journal articles/book chapters and conference papers. Mr. Subrata received a fellowship from the Information and Communication Technology Division of the Ministry of Posts, Telecommunications, and Information Technology of Bangladesh. Mr. Subrata is also a Graduate member of IEEE.



Rakhi Rani Paul graduated with her B.Sc. and M.Sc. Engineering from Rajshahi University, in Computer Science and Engineering in 2017 and 2018, respectively. Currently, she works as a Lecturer at the Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad Cantonment, Natore-6431. Bangladesh. He has published more than 20 international journal articles/book chapters and conference papers. Her research fields are Deep Learning, Data Mining, and Speech Signal Processing.



Md Ekramul Hamid received the B.Sc. and M.Sc. degrees in Applied Physics and Electronics from the University of Rajshahi. Later, I received a Master's degree in Computer Science from the University of Pune, India, and a PhD 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 60 international journal/conference papers. He is a recipient of the Monbukagakusho Scholarship, JASSO Fellowship, and NIST Fellowship for his contributions 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. He worked as the Chairman of the CSE department and the Dean of the Faculty of Engineering at the University of Rajshahi. His research interests include Digital signal processing, Machine learning, Analysis and synthesis of speech signals, Speech enhancement, and Image processing.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.