An Efficient Patient Activity Recognition using LSTM Network and High-Fidelity Body Pose Tracking

Thanh-Nghi Doan
Faculty of Information Technology, An Giang University
Vietnam National University Ho Chi Minh City
An Giang, Vietnam

***Abstract*—The need for healthcare services is growing, particularly in light of the COVID-19 epidemic's convoluted trajectory. This causes overcrowding in medical facilities, making it difficult to manage, treat, and monitor patients' health. Therefore, a method to remotely observe the patient's behavior is required, to aid in early warning and treatment, and to reduce the need for hospitalization for patients with minor diseases. This paper proposes a new real-time smart camera system to monitor, recognize and warn the patient's abnormal actions remotely with reasonable cost and easy to deploy in practice. The key benefit of the proposed methods is that patient actions may be detected without the usage of ambient sensors by employing pictures from a regular video camera. It carries out the detection using high-fidelity human body pose tracking with MediaPipe Pose. Then, the Raspberry Pi 4 device and the LSTM network are used for remote monitoring and real-time classification of patient actions. The test dataset is built from reality and reuses the existing datasets. Our system has been evaluated and tested in practice with over 96.84% accuracy, runs at over 30 frames per second, suitable for real-time execution on mobile devices with limited hardware configuration.**

***Keywords*—Human body pose tracking; LSTM; raspberry Pi 4; patient monitoring system**

I. Introduction

The development of efficient and reliable remote patient action recognition systems has been receiving much attention from the scientific research community. The benefits of patient monitoring from a distance include the ability to detect illnesses early and in real time, monitor patients continuously, stop illnesses from getting worse and prevent untimely deaths, lower hospitalization costs, fewer hospitalizations, and more accurate readings while still allowing patients to go about their daily lives normally. By using communication technology, emergency medical services, care for patients with mobility issues, emergency care for injuries sustained in traffic accidents and other types of accidents, and non-invasive medical interventions, healthcare services are made more efficient. In recent years, action recognition methods have focused heavily on the use of image and video analysis technologies. There are different definitions of action recognition presented in the study by Herath et al. [1]. The rapid development of smart devices and deep learning techniques have spurred the development of action recognition systems. These techniques have been widely applied in life such as entertainment, monitoring and human health care [2]. However, according to the survey by Szegedy et al. [3], the identification of complex and specific actions is still a big challenge to study. The articles [4], [5] presented a comprehensive review of fall detection systems and remote patient action recognition. Researchers have built a large variety of systems that can operate with the many technologies used to monitor patient behavior. These systems are broadly characterized as wearable, ambient, and computer-vision-based [6]. Researchers have created a vast array of systems that can work with the various technologies that are used to track patient activity. In general, these systems can be divided into wearable, ambient, and computer-vision-based ones [6].

The first block, wearable systems, includes sensors carried by the monitored individual. This set of systems employs a wide range of technologies, including accelerometers, pressure sensors, inclinometers, gyroscopes, and microphones, among others. The authors of [7] carefully evaluate and study these systems. The study attempts to assess the state of the art in such monitoring, both in terms of the most commonly used sensor technologies and their placement on the human body. These techniques, however, have the problem of requiring the devices to be put on the individuals' bodies. Because this sort of sensor must be worn continually, it might be unpleasant and not always viable [8]. The second block contains devices with pressure, acoustic, infrared, and radio-frequency sensors that are positioned around the monitored individual [9], [10]. However, the expense of installing these systems is very costly, and they are only appropriate for specialist patient care rooms, making them difficult to employ in everyday living at home. The last block, which is the focus of this research, groups systems capable of identifying human recognition using image-based computer vision. In recent years, convolutional neural networks (CNN) [11] have been widely used in image classification problems in many fields. Due to CNN's superior performance [12], many studies have started to use CNN for video classification. Long Short-Term Memory (LSTM) neural networks and conventional CNNs, or a combination of the two, have both been shown to perform well in human action recognition (HAR). In which CNN has been used to analyze sensor data for HAR with exceptional results [13]. Previous studies have proposed to supplement the feature vector extracted by CNN with some statistical features [14]. Aviléz-Cruz et al. [15] have developed a three-input CNN model to

recognize six human actions. The usefulness of LSTM networks for HAR has been demonstrated by additional studies as well [16]. Finally, several studies have suggested enhancing CNN with LSTM layers [17]. Recently, the article [18] proposed a network model that combines LSTM, MobileNetV2 and Raspberry Pi 4 in remote patient action monitoring and identification. However, these methods in the last block have the drawback of making it difficult to distinguish between closely related actions, such as waving and clapping, walking and running. Furthermore, training network models takes a long time due to direct learning of data from video frames, where CNN models are used to extract features from video frames. Because of this, these methods require extensive hardware configurations and have slow response times.

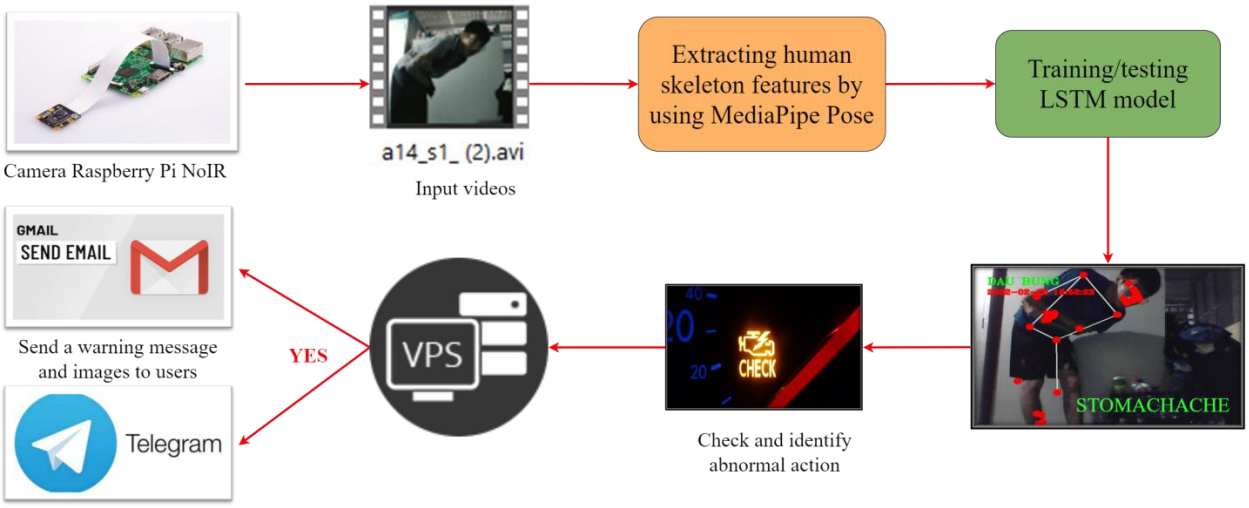

Fig. 1. Overview of our Proposed System for Monitoring Patient Healthy.

On the other hand, studies on the skeleton features extraction method for patient action identification [19], [20], and [21] have demonstrated a number of benefits that can get around the issues raised above. As a result, this paper proposes a new system for real-time identification of patient actions that is low in cost, efficient, has a fast response time, and is simple to install and implement in practice using hardware devices with limited configuration. The main contributions of the paper include:

- A novel method for patient action recognition using MediaPipe Pose framework [22], LSTM network model and Raspberry Pi 4 device [23].
- A new dataset was created using 541 videos of 16 different types of patient actions that were preprocessed and labeled in accordance with benchmark dataset standards.
- An action recognition system that is fully developed, user-friendly, and capable of continuous monitoring and alerting of abnormal human activity of the patient at home.

The rest of the article is arranged as follows. Section II describes the materials and methods used to describe overview of our system, data collection, patient action recognition model, Raspberry Pi and Camera Raspberry Pi NoIR . The experimental results and discussion are reported in Section III. Section IV presents the conclusions, limitations, and recommendations for future research.

II. Materials and Methods

A. Overview of our System

The overview of the proposed system for the remote patient monitoring camera system is shown in Fig. 1. In which the Raspberry Pi 4 Camera Module NoIR [23] is used to continuously monitor the patient's activities at home in real time. The generated video sequence is recognized and labeled in real time using the MediaPipe Pose framework [24] and the LSTM network model [25] trained and saved on a Raspberry Pi 4 device. If the patient's actions are considered to be abnormal, there is a health problem, and the system will immediately send a warning message, along with a photo of the odd activity, to the patient's relatives via email and the *Telegram* messaging application. The labeled videos are then saved on a virtual server on a regular schedule. Videos labeled as abnormal actions will be kept on the server for a long time, whereas normal actions will be kept for a short time and deleted after a certain period of time to save storage space. The algorithm to recognize real-time patient actions in videos and send warning messages with images of abnormal actions to users is summarized and shown in Fig. 2.

B. Data Collection

There are many published datasets on human action recognition such as ActivityNet [26], Kinetics [27], UCF101 [28], HMDB51 [29], STAIR-Actions [30], KARD [31], and NTU RGB+D [32]. However, these datasets do not include recordings of patient activities, and there is presently no published official benchmark dataset for these types of activities. Therefore, this study self-constructed a new dataset on patient actions to test our proposed approach. This dataset combines existing data with data generated by us from the actual world. A summary of this dataset is presented in Table I.

```
BEGIN
    Input: LSTM model, patient body skeleton points
    Label = "Nomarl action"
    The model predict the result based on the skeleton points
    IF Result = 1:
        Label = 'Hand swing'
    ELIF Result =2:
        Label = 'Hand clap'
     .
     .
     .
    ELIF Result =14:
    {
        Label = 'Stomachache'; Save the stomachache images
        Send emails, messages, images via Gmail and Telegram
    }
     .
     .
     .
    ELSE
        Label = 'Normal action'
    RETURN: Label
END
```

Fig. 2. The Algorithm Processes Skeleton Point Data, Returns Results and Sends Notification Messages to GMAIL, TELEGRAM.

TABLE I. A SUMMARY OF OUR DATASET ITH 16 TYPES OF PATIENT ACTIONS

ID	Description	Avg frame	Frame		Frame/s	Number of videos
			Width	Height		
a01	Hand swing	2500	640	480	25	81
a02	Hand clap	2500	640	480	25	23
a03	Body swing	2500	640	480	25	24
a04	Drink	2500	640	480	25	51
a05	Sit down	2500	640	480	25	34
a06	Stand up	2500	640	480	25	33
a07	Walking	2500	640	480	25	55
a08	Side kick	2500	640	480	25	66
a09	Phone call	2500	640	480	25	34
a10	Hand pain	2500	640	480	25	15
a11	Leg pain	2500	640	480	25	22
a12	Headache	2500	640	480	25	25
a13	Neck pain	2500	640	480	25	25
a14	Stomachache	2500	640	480	25	22
a15	Backache	2500	640	480	25	19
a16	Fall down	2500	640	480	25	12
Total number of videos						**541**

Due to time and staffing restrictions, we could only create a test dataset with 16 examples of the patient's actions. These actions are collected and separated into two groups: (i) the patient's normal actions (shown in the blue bounding box of Fig. 3) and (ii) the patient's abnormal actions (shown in the red bounding box of Fig. 3. This dataset includes four actions taken from the KARD dataset [31] and 12 actions we independently created by recording patient action video clips in the real experimental setting. KARD is a dataset that includes 18 different types of indoor daily activities with a resolution of 640x480 and reasonably clear action gestures. Consequently, they can be utilized to develop and evaluate a patient health monitoring system at home. However, only four action classes that are appropriate for this problem are used in this study: sit down, stand up, side kick, and phone call. We recruited volunteers to carry out 12 distinct types of actions for fact-generated data. Each type of action was performed three times and video recorded for three seconds each, using a Webcam HD 720p with the detailed specifications shown in Table II.

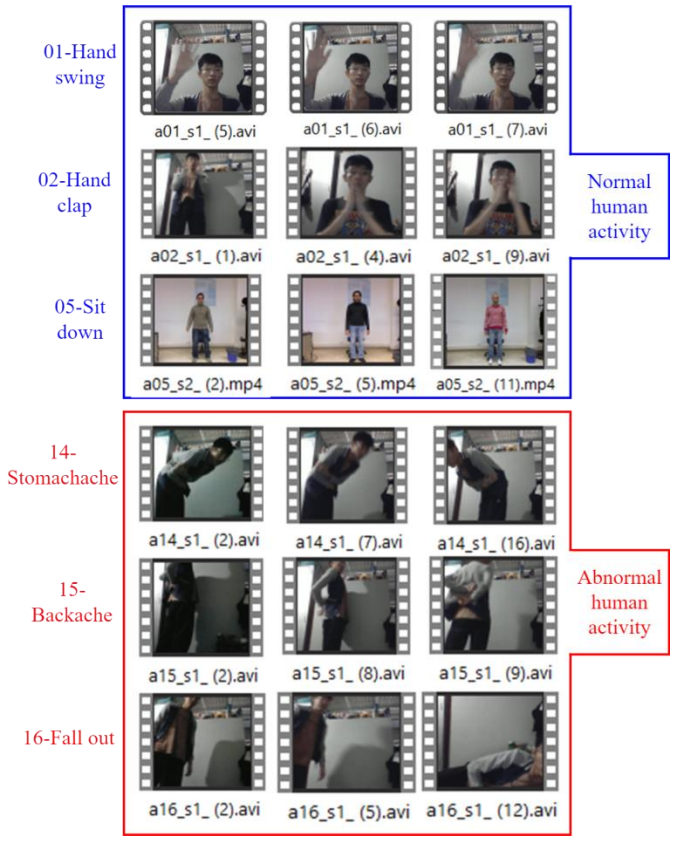

Fig. 3. Video Samples of our Patient Activities Datasets.

TABLE II. THE DETAILED SPECIFICATIONS OF WEBCAM HD 720P.

Specifications	Value
Camera	HD webcam 720p
Resolution	800x600
FPS	25 frames/second
Camera color	Color
Flash mode	No
Focus type	Fixed focus
Video format	AVI

The total number of videos we have collected is over 700 videos. These videos are then preprocessed, and the videos that do not meet the quality requirements are removed, yielding a video dataset of 541 files. The total size of the video dataset is 241 MB in which, the number of videos of each action type ranges from 12 to 81 videos, as shown in Table I. Each video is

shot at a frame rate of 25 FPS. This dataset is annotated in order of videos in each folder according to each action type, i.e. a01_s1_(1).mp4, a01_s01_(2).mp4,…, a02_s1_(1).mp4, a02_s01_(2).mp4,… equivalent to actions labeled as a01 (Hand swing), a02 (Hand clap), a03 (Body swing), a04 (Drink), a05 (Sit down), a06 (Stand up), a07 (Walking)), a08 (Side kick), a09 (Phone call), a10 (Hand pain), a11 (Leg pain), a12 (Headache), a13 (Neck pain), a14 (Stomachache), a15 (Backache) and a16 (Fall down). Some sample videos of the dataset consisting of 09 normal actions and 07 abnormal actions of the patient, are presented as shown in Fig. 3.

C. Patient Action Recognition Model

1) Extract Human Body Features with MediaPipe Pose:

MediaPipe Pose is a machine learning solution for high-fidelity body pose monitoring that uses the BlazePose research [22] to infer 33 3D landmarks and a background segmentation mask on the entire body from RGB video frames. The network can generate 33 body keypoints for a single human during inference and performs at over 30 frames per second on a Pixel 2 phone. Therefore, it is well suited to real-time applications such as fitness tracking and sign language recognition. The benefit of this skeletal feature extraction method is its real-time speed, fast response time, and good results even with low-quality and low-resolution video clips, independent of ambient variables such as light, shadow, and the ability to identify many objects at the same time.

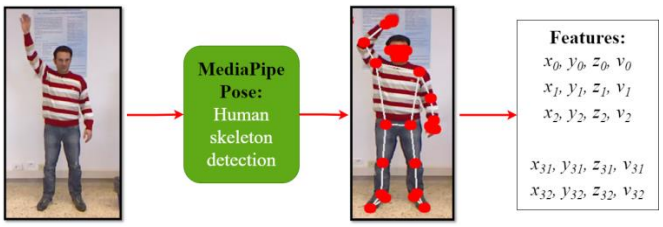

Fig. 4. Extracting Human Skeleton Features using MediaPipe Pose [24].

In this study, skeleton features are extracted from each frame of video using the algorithm described in the paper [22]. As shown in Fig. 4, the result of extracting each video frame is 33 skeleton points corresponding to 33 coordinates (*x*, *y*, *z*) and a visibility value *v* numbered from 0 to 32. Each skeleton point is assigned a different ID number when stored in the file used for model training. Since the video is recorded at a frame rate of 25 FPS, the number of captured frames is calculated as 25 x video recording time in second. As a result, the total number of frames collected from the dataset of 16 action classes is 19,345 frames, as shown in Table III.

2) LSTMs for Patient Activity Recognition:

Long Short-Term Memory [25] is a Recurrent Neural Network (RNN) that has been increasingly used in the field of deep learning and human action recognition. LSTM, as opposed to standard feedforward neural networks, includes feedback connections. A recurrent neural network of this type can analyze not just individual data points (such as photographs), but also whole data sequences (such as speech or video). For instance, LSTM may be used for handwriting recognition, speech recognition, and anomaly detection in network traffic or intrusion detection systems. A typical LSTM unit comprises of a cell, an input port, an output port, and a forget port (shown in Fig. 5). The cell stores values for an indefinite amount of time, and the three gates control the flow of information into and out of the cell.

LSTM networks are well suited for classification, processing, and prediction based on time series data because they can handle indeterminate delays between significant events in time series. LSTM was developed to solve the vanishing gradient problem that can be encountered when training traditional RNNs. The benefit of LSTM over standard RNNs, Hidden Markov models, and other sequential learning approaches is its low sensitivity across a particular length range. RNNs can, in theory, follow any long-term relationships in input sequences. The problem with RNNs, however, is computational nature: when training an RNN using backpropagation, the backpropagation gradients can be degraded (i.e. tend to move towards zero) or "explode" towards infinity. Because LSTM units allow gradients to remain constant, RNNs utilizing LSTM units can partially alleviate the gradient degradation problem. However, these LSTMs can still suffer from gradient "explosion" problems.

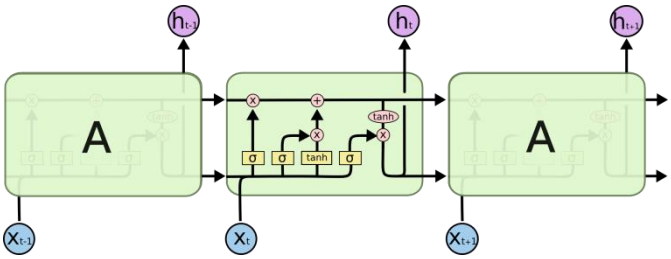

Fig. 5. Illustration of an LSTM Cell and Architecture. The Repeater Module in an LSTM Contains Four Interaction Layers.

In this work, a multi-layer LSTM with four layers have been implemented for patient activity recognition. Each layer has 50 units and is followed by a dropout layer designed to decrease the model's overfitting to the training data. Finally, a dense fully connected layer with 27 units is utilized to interpret the features retrieved by the LSTM hidden layer before making predictions with a final output layer with softmax function. The efficient Adam version of stochastic gradient descent will be utilized to optimize the network, and the categorical cross entropy loss function will be used because we are learning a multi-class classification problem.

D. Raspberry Pi and Camera Raspberry Pi NoIR

The Raspberry Pi [23] is a tiny computer developed by the Raspberry Pi Foundation in collaboration with Broadcom in the United Kingdom. The original Raspberry Pi project's goal was to promote basic computer literacy education in schools and developing countries. This device, however, became unexpectedly popular and was marketed for the purpose of building robots. Because of its inexpensive cost and open design, it is frequently utilized in various sectors, including weather monitoring. After the second version was released, the Raspberry Pi Foundation produced a brand-new gadget called the Raspberry Pi Trading. Raspberry Pi 4 Model B was released in June 2019 [33] with a 1.5 GHz quad-core ARM Cortex-A72 processor, 802.11ac Wi-Fi, Bluetooth 5, gigabit Ethernet (unlimited throughput), two USB 2.0 ports, two USB 3.0 ports, 2-8 GB RAM, and dual monitor support via a pair of

micro HDMI ports for up to 4K resolution. When used in conjunction with an appropriate power supply, the Raspberry Pi 4 is also powered via a USB-C port, allowing additional power to be provided to downstream peripherals (Fig. 6).

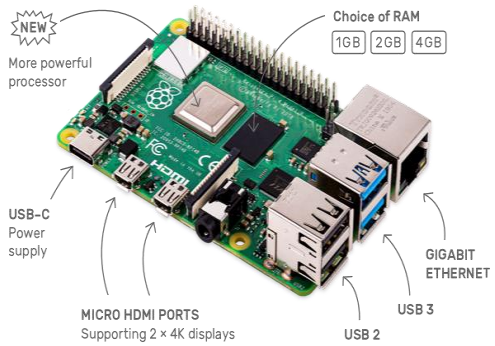

Fig. 6. Raspberry Pi 4 Model B.

The Raspberry Pi NoIR V2 IMX219 Camera (Fig. 7) is the latest version of the Camera Module for Raspberry Pi that uses the 8-megapixel IMX219 image sensor from Sony instead of the old OV5647 sensor. With the 8-megapixel IMX219 sensor from Sony, the Camera Module for Raspberry Pi has achieved a remarkable upgrade in both image and video quality as well as durability. Raspberry Pi NoIR V2 IMX219 8 MP camera can be used with Raspberry Pi to take photos and videos in low light conditions with HD 1080p30, 720p60, or VGA90 quality. It's also very simple, as we only need to connect the Raspberry Pi's Camera port and config to run the program. The Raspberry Pi NoIR V2 IMX219 8 MP camera is controllable via MMAL and V4L APIs, there are many libraries developed by the Raspberry Pi community on Python that make learning and using it much easier.

Fig. 7. Camera Raspberry Pi NoIR V2.

III. Experimental Results and Discussion

A. Experimental Settings and Evaluation Metric

All experiments were conducted on the laptop Asus TUF Gaming FX506LH i5 10300H, 8 GB RAM, NVIDIA GeForce GTX 1650 4GB, with the Ubuntu operating system. The real-time recognition system's algorithm is written in Python and implemented on the Linux operating system using the open-source libraries Keras and OpenCV (Raspberry Pi OS). Accuracy is a metric used to evaluate how well classification models perform and is calculated using the formula (1).

$$Accuracy = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples}} \quad (1)$$

B. Model Training and Evaluation

The dataset of 16 action classes was split into two subsets that were used to train and evaluate the model at an 80:20 ratio, as was done by Luhach et al. [34]. Thus, the dataset has 15,476 frames for training and the remaining 3,869 frames for model evaluation. The system uses the data converted from video frames to coordinates (x, y, z) and the visibility value of the skeleton point features to train the model when using MediaPipe Pose. All this data is stored in CSV format file. Dataset for training and evaluating models are described in Table III.

TABLE III. Dataset Information for Model Training and Evaluation

Number of samples	Training	Testing	Timesteps	Dimension
19,345 frames	19,345 × 0.8 = 15,476 frames	19,345 × 0.2 = 3,869 frames	20	132 (33 points x 4 values) × Time-steps × number of frames

Since elaborate hyperparameter optimization methods like grid search were judged too time-consuming for the scope of this study, the various parameter-settings for the training process were developed via trial and error. As a result, many parameter choices have been attempted and tested by repeatedly running the model, and values for the hyperparameters that are thought to be near to an equilibrium between time-efficiency and performance have been chosen. In the selection process, several settings that reduced the difference between high and low values were tested, which is not unlike to how many root finding techniques in mathematics operate.

Different batch sizes (number of samples per gradient update) were examined, and 32 was found to be an appropriate value in terms of both effectiveness and performance. It was decided to iterate through the full dataset 50 times because epoch sizes greater than 50 produced negligible to no improvements. The mean squared error is employed for loss function. Adam produced the best results of the several optimizers available and was hence chosen over stochastic gradient descent. Finally, in the LSTM network model, the time-steps K are the most critical parameters affecting model performance. The time-steps are how many lagged variables the model receives as input to forecast the following step. Therefore, various number of time-steps are examined to determine how they impact the model's performance. The time-steps chosen to be tested are 5, 10, 15, and 20. The model is trained and evaluated five times for each of these time-steps in order to gather sufficient data to compare their relative performance. The performance of the model using different time-steps is illustrated in Table IV.

TABLE IV. The Overall Performance of the Model with Different Time-Steps.

K	5	10	15	20
Accuracy	92.26%	95.63%	96.44%	96.84%
Loss	0.1838	0.1047	0.0914	0.0854

Table IV shows that as the number of time-steps *K* is increased, the model's performance improves (accuracy increases and loss lowers), but training time increases. For instance, a model with 10 time-steps segments performs better than one with 5 time-steps (95.63 percent vs 92.26 percent). This result demonstrates that the model will learn more information from earlier frames if many time-steps *K* are used. As a result, the resulting features of the videos will be more robust and high-abstract, improving the model's classification precision. However, when *K* is increased to 20, the model performance exhibits evidence of saturation at 96.44% as opposed to 96.44% with *K* = 15. Therefore, we set *K* equal to 20 for the model to achieve the best classification performance while keeping training and evaluation time to a minimum. Fig. 8 depicts the curve reflecting the model's accuracy and loss after 50 iterations.

After training, the resulting model size is only 1.2 MB, making it appropriate for installation on Raspberry Pi devices with limited memory configuration. According to the article [22], the FPS of BlazePose Full is 102, while that of BlazePose Lite is 312, making it ideal for developing real-time applications. The model training process is quite fast, averaging about 5–10 seconds for an iteration with 16 action classes, because the video data has been converted to a text file in CSV format, so it doesn't take a lot of hardware resources. The model is trained in 50 iterations taking from 5 to 10 minutes. The resulting model has achieved an accuracy of 96.84% on our dataset. After the real-time action recognition model's training and evaluation on the Rapberry Pi 4 system produced good and consistent results, the system was installed and tested to send alert messages and emails if the camera detects unusual patient health-related behaviors (shown as shown in Fig. 9 and Fig. 10). Fig. 11 depicts our system's successful detection of six real-time patient actions: hand clap, sit down, stomachache, backache, and fall down. Our next step is to collect more data from a variety of patient actions and then to investigate mobile devices with better hardware configurations, such as the Jetson Nano Developer Kit [35] and CNN models that are efficient, accurate, and suitable for the latest mobile devices.

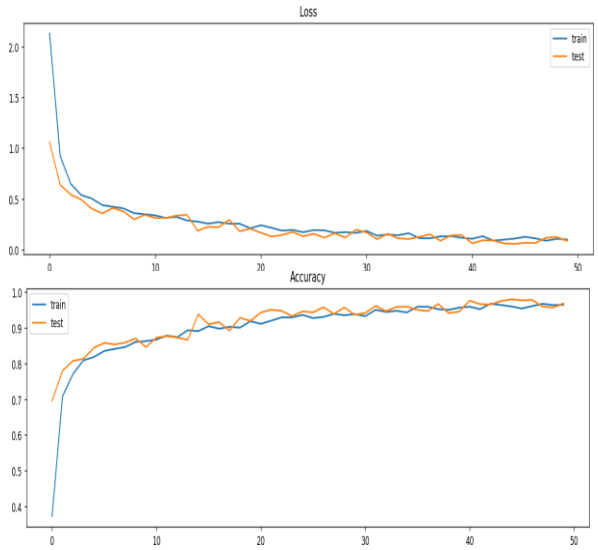

Fig. 8. Accuracy and Loss of Training Process with 50 Epoches.

Fig. 9. Patient Alert Emails are sent to the Users.

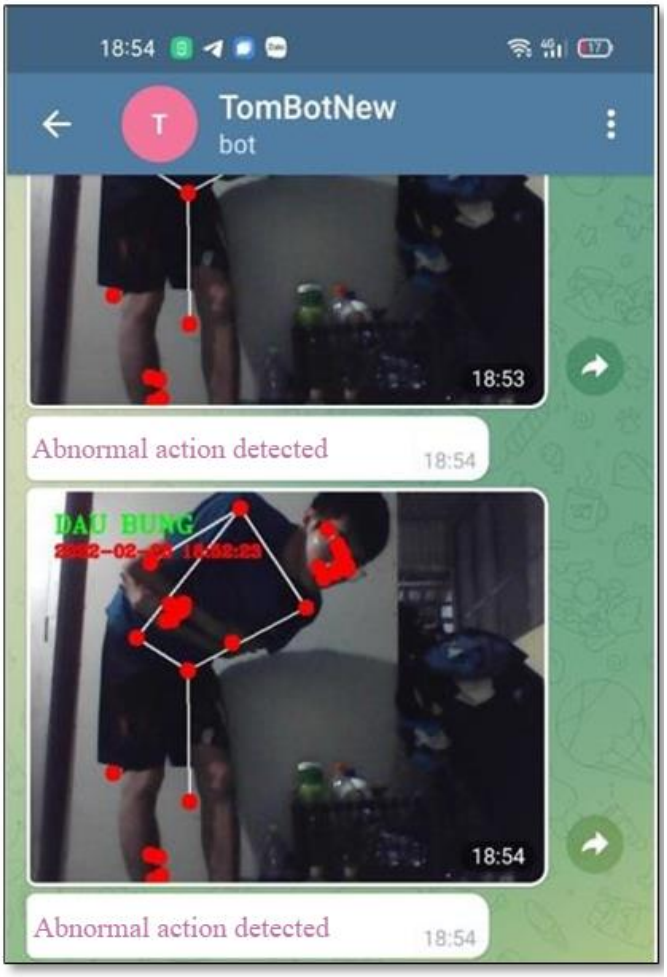

Fig. 10. Alert Messages and Images are sent to the Users via the Telegram Application.

Fig. 11. Real Time Patient Action Identification Results on Raspberry pi 4, Camera Module Noir System.

IV. Conclusion and Future Research Work

This study has proposed a novel system with basic functions of a smart surveillance camera, supporting remote patient monitoring. A model for remote skeletal patient activity detection was developed using MediaPipe Pose, an LSTM network, and a Raspberry Pi 4. The numerical results show that our proposed model performed well in classification, with an accuracy of 96.84% on a dataset of 16 activities gathered and constructed by ourselves. In addition, because the MediaPipe Pose library and the LSTM network are used for recognition, the recognition model size is small, and the network training parameters are few, making it appropriate for deployment on mobile devices with limited hardware, such as the Raspberry Pi 4. Therefore, our method offers numerous benefits in terms of real-time patient action recognition, low cost, simple installation, and practical implementation. A dataset of 541 video files of patients' actions in indoor was built to evaluate our method. Although the amount of data is little and there isn't much actual patient data, this provides the foundation for future larger, better-quality data sets that will help the research community better understand patient activities.

References

[1] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," Image Vis. Comput., vol. 60, pp. 4–21, 2017, doi: 10.1016/j.imavis.2017.01.010.

[2] Y. Bengio, "Deep Learning of Representations: Looking Forward," ArXiv, vol. abs/1305.0, 2013.

[3] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, no. January 2017, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.

[4] J. Gutiérrez, V. Rodríguez, and S. Martin, "Comprehensive review of vision-based fall detection systems," Sensors (Switzerland), vol. 21, no. 3, pp. 1–50, 2021, doi: 10.3390/s21030947.

[5] L. P. Malasinghe, N. Ramzan, and K. Dahal, "Remote patient monitoring: a comprehensive study," J. Ambient Intell. Humaniz.

Comput., vol. 10, no. 1, pp. 57–76, 2019, doi: 10.1007/s12652-017-0598-x.

[6] P. Vallabh and R. Malekian, “Fall detection monitoring systems: a comprehensive review,” J. Ambient Intell. Humaniz. Comput., vol. 9, no. 6, pp. 1809–1833, 2018, doi: 10.1007/s12652-017-0592-3.

[7] R. Rucco et al., “Type and location of wearable sensors for monitoring falls during static and dynamic tasks in healthy elderly: A review,” Sensors (Switzerland), vol. 18, no. 5, 2018, doi: 10.3390/s18051613.

[8] Z. Liu, Y. Cao, L. Cui, J. Song, and G. Zhao, “A Benchmark Database and Baseline Evaluation for Fall Detection Based on Wearable Sensors for the Internet of Medical Things Platform,” IEEE Access, vol. 6, pp. 51286–51296, 2018, doi: 10.1109/ACCESS.2018.2869833.

[9] S. Cheng, L. Thomas, J. Cook, and M. Pecht, “A Radio Frequency Sensor System for Prognostics and Health Management,” 2009, doi: 10.1115/DETC2009-87723.

[10] M. Jang, S. Kang, and S. Lee, “Monitoring Person on Bed Using Millimeter-Wave Radar Sensor,” in 2022 IEEE Radar Conference (RadarConf22), 2022, pp. 1–4, doi: 10.1109/RadarConf2248738.2022.9764251.

[11] Q. Li, W. Cai, X. Wang, Y. Zhou, D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” 2014 13th Int. Conf. Control Autom. Robot. \& Vis., pp. 844–848, 2014.

[12] J. Gu et al., “Recent advances in convolutional neural networks,” Pattern Recognit., vol. 77, no. June 2016, pp. 354–377, 2018, doi: 10.1016/j.patcog.2017.10.013.

[13] M. Z. Uddin and M. Hassan, “Activity Recognition for Cognitive Assistance Using Body Sensors Data and Deep Convolutional Neural Network,” IEEE Sens. J., vol. 19, pp. 8413–8419, 2019.

[14] A. D. Ignatov, “Real-time human activity recognition from accelerometer data using Convolutional Neural Networks,” Appl. Soft Comput., vol. 62, pp. 915–922, 2018.

[15] C. Avilés-Cruz, A. Ferreyra-Ramírez, A. Zúñiga-López, and J. Villegas-Cortéz, “Coarse-fine convolutional deep-learning strategy for human activity recognition,” Sensors (Switzerland), vol. 19, no. 7, 2019, doi: 10.3390/s19071556.

[16] W.-H. Chen, C. Baca, and C.-H. Tou, “LSTM-RNNs combined with scene information for human activity recognition,” 2017 IEEE 19th Int. Conf. e-Health Networking, Appl. Serv., pp. 1–6, 2017.

[17] H. Li and M. Trocan, “Personal Health Indicators by Deep Learning of Smart Phone Sensor Data,” 2017 3rd IEEE Int. Conf. Cybern., pp. 1–5, 2017.

[18] T. Thanh-Nghi, D., Thanh-Hien-Triet, N., Truong-An, “Smart camera system for remote patient activity monitoring,” in the 14th National Scientific Conference on Research and Application of Information Technology - Fair’ 2021, Natural Science and Technology Publishing House, 2021, pp. 110–117.

[19] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, “Fall Detection and Activity Recognition Using Human Skeleton Features,” IEEE Access, vol. 9, pp. 33532–33542, 2021, doi: 10.1109/ACCESS.2021.3061626.

[20] J.-C. Chiang et al., “Posture Monitoring for Health Care of Bedridden Elderly Patients Using 3D Human Skeleton Analysis via Machine Learning Approach,” Appl. Sci., vol. 12, no. 6, p. 3087, 2022, doi: 10.3390/app12063087.

[21] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, “Skeleton-based human activity recognition using ConvLSTM and guided feature learning,” Soft Comput., vol. 26, no. 2, pp. 877–890, 2022, doi: 10.1007/s00500-021-06238-7.

[22] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device Real-time Body Pose tracking,” CoRR, vol. abs/2006.1, 2020, [Online]. Available: https://arxiv.org/abs/2006.10204.

[23] W. Gay, Raspberry Pi Hardware Reference. Apress Berkeley, CA, 2014.

[24] C. Lugaresi et al., “MediaPipe: A Framework for Building Perception Pipelines,” 2019, [Online]. Available: http://arxiv.org/abs/1906.08172.

[25] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[26] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 07-12-June. pp. 961–970, 2015, doi: 10.1109/CVPR.2015.7298698.

[27] W. Kay et al., “The Kinetics Human Action Video Dataset,” ArXiv, 2017, [Online]. Available: http://arxiv.org/abs/1705.06950.

[28] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” no. December 2012, 2012, [Online]. Available: http://arxiv.org/abs/1212.0402.

[29] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” Proc. IEEE Int. Conf. Comput. Vis., pp. 2556–2563, 2011, doi: 10.1109/ICCV.2011.6126543.

[30] Y. Yoshikawa, J. Lin, and A. Takeuchi, “STAIR Actions: A Video Dataset of Everyday Home Actions.” 2018, [Online]. Available: http://arxiv.org/abs/1804.04326.

[31] S. Gaglio, G. Lo Re, and M. Morana, “Human Activity Recognition Process Using 3-D Posture Data,” IEEE Trans. Human-Machine Syst., vol. 45, no. 5, pp. 586–597, 2015, doi: 10.1109/THMS.2014.2377111.

[32] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem. pp. 1010–1019, 2016, doi: 10.1109/CVPR.2016.115.

[33] E. Upton, “Raspberry Pi 4 on sale now from $35.” Raspberry Pi Foundation, 2019.

[34] A. Luhach, D. Jat, K. Ghazali, X.-Z. Gao, and P. Lingras, Advanced Informatics for Computing Research: Third International Conference, ICAICR 2019, Shimla, India, 15–16 June 2019; Revised Selected Papers; Springer Nature: Berlin/Heidelberg, Germany, 2019; Volume 1075. 35. 2019.

[35] NVIDIA, “Jetson Nano Developer Kit User Guide.” 2021, [Online]. Available: https://developer.download.nvidia.com/ embedded/L4T/r32-3-1_Release_v1.0/Jetson_Nano_ Developer_Kit_User_Guide.pdf.