Automatic Assessment of Upper Extremity Function and Mobile Application for Self-Administered Stroke Rehabilitation

Dong-Wook Kim, Ji Eun Park, Min-Jung Kim, Seung Hwan Byun, Chung In Jung, Ha Mok Jeong, Sang Rok Woo, Kwon Haeng Lee, Myoung Hwa Lee, Jung-Woo Jung, Dayeon Lee, Byung-Ju Ryu, Seung Nam Yang, and Seung Jun Baek

***Abstract*— Rehabilitation training is essential for a successful recovery of upper extremity function after stroke. Training programs are typically conducted in hospitals or rehabilitation centers, supervised by specialized medical professionals. However, frequent visits to hospitals can be burdensome for stroke patients with limited mobility. We consider a self-administered rehabilitation system based on a mobile application in which patients can periodically upload videos of themselves performing reach-to-grasp tasks to receive recommendations for self-managed exercises or progress reports. Sensing equipment aside from cameras is typically unavailable in the home environment. A key contribution of our work is to propose a deep learning-based assessment model trained only with video data. As all patients carry out identical tasks, a fine-grained assessment of task execution is required. Our model addresses this difficulty by learning RGB and optical flow data in a complementary manner. The correlation between the RGB and optical flow data is captured by a novel module for modality fusion using cross-attention with Transformers. Experiments showed that our model achieved higher accuracy in movement assessment than existing methods for action recognition. Based on the assessment model, we developed a patient-centered, solution-based mobile application for upper extremity exercises for hemiplegia, which can recommend 57 exercises with three levels of difficulty. A prototype of our application was evaluated by potential end-users and achieved a good quality score on the Mobile Application Rating Scale (MARS).**

***Index Terms*— Deep learning, hemiplegia, motion assessment, self-administered rehabilitation, upper extremity.**

I. Introduction

UPPER limb motor impairment is common and has been reported in more than 80% of patients after stroke. Less than half of the patients regain basic functions of the upper limb by 12 months, by recovering from disabilities which markedly restrict their independence in activities of daily living [1], [2]. Patients with motor impairment due to stroke experience significant limitations in their daily lives. Thus, motor relearning programs with regular and repetitive training are important for the successful recovery of stroke patients with impaired movement [3], [4], [5]. Most rehabilitation programs require supervision and guidance from experts and are conducted in hospitals or rehabilitation centers. However, for stroke patients with limited mobility, consecutive hospital visits for rehabilitation and treatment are a burden, which makes it challenging to accurately evaluate their individual motor function or select appropriate exercises. Thus, *self-administered rehabilitation* is deemed to facilitate patient recovery, because patients show greater willingness to exercise frequently and repetitively, which can promote neuroplasticity [6], [7], [8], than to visit the hospital. The rehabilitation with motor training should be based on an accurate evaluation of individual motor function, because the upper limb dysfunction varies among patients. Therefore, the development of a

Manuscript received 14 August 2023; revised 21 December 2023 and 10 January 2024; accepted 20 January 2024. Date of publication 25 January 2024; date of current version 1 February 2024. This work was supported in part by the ICT Creative Consilience Program supervised by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2020-0-01819; in part by the National Research Foundation of Korea (NRF) Grant by the Ministry of Science and ICT (MSIT) under Grant 2021R1A2C1007215 and Grant 2022R1A5A1027646; in part by the Korea Medical Device Development Fund Grant by MSIT under Project 1711195279 and Project RS-2021-KD000009; in part by the Ministry of Trade, Industry and Energy; in part by the Ministry of Health and Welfare; in part by the Ministry of Food and Drug Safety of the Korean Government; and in part by the Basic Science Research Program through NRF funded by the Ministry of Science, ICT and Future Planning, under Grant NRF-2019R1G1A1009646. *(Dong-Wook Kim and Ji Eun Park are co-first authors.) (Corresponding authors: Seung Jun Baek; Seung Nam Yang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Regional Institutional Review Board of the Korea University Guro Hospital under IRB No. 2021GR0178.

Dong-Wook Kim, Min-Jung Kim, Seung Hwan Byun, Chung In Jung, and Seung Jun Baek are with the Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: tigerrhs@korea.ac.kr; kmj0606@korea.ac.kr; senghan1992@korea.ac.kr; whooznext@korea.ac.kr; sjbaek@korea.ac.kr).

Ji Eun Park, Ha Mok Jeong, Sang Rok Woo, Kwon Haeng Lee, Myoung Hwa Lee, Jung-Woo Jung, Dayeon Lee, and Seung Nam Yang are with the Department of Physical Medicine and Rehabilitation, Korea University Guro Hospital, Korea University College of Medicine, Seoul 08308, Republic of Korea (e-mail: micky1322@naver.com; jhmgkahr@naver.com; mission365@naver.com; snap9@hanmail.net; tinimung@naver.com; tesejw382@gmail.com; ekdus58@korea.ac.kr; snamyang@korea.ac.kr).

Byung-Ju Ryu is with the Department of Physical Medicine and Rehabilitation, Sahmyook Medical Center, Seoul 02500, Republic of Korea (e-mail: btjrbj@gmail.com).

Digital Object Identifier 10.1109/TNSRE.2024.3358497

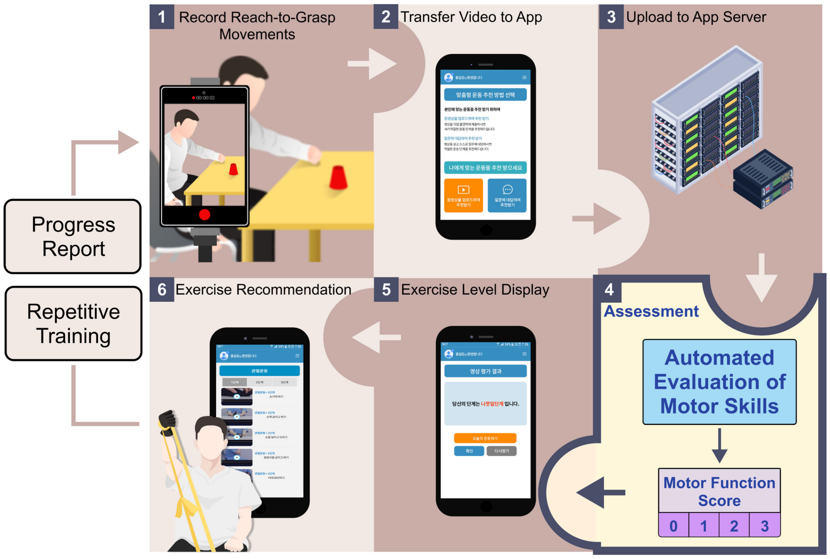


Fig. 1. Overview of a self-administered rehabilitation system for stroke survivors. There are six steps in total, and each step is described as follows: 1) Record a video of a patient executing reach-to-grasp movements. 2) Transfer the recorded video to the application. 3) Upload the video to the server for assessment. 4) Automatically evaluate the score of motor function on a scale of 0 to 3. 5) Display the patient's exercise level based on the score.6) Recommend appropriate exercises according to the patient's level. The key component of our system is Step (4): Automated assessment of motor function.

self-rehabilitation program with accurate evaluation capabilities will greatly benefit the recovery of stroke patients.

In a self-administered exercise programme, it is important to match the exercise level to the patient's upper limb disability. Thus, we consider the framework of motion assessment followed by exercise recommendation. Specifically, we propose a self-administered rehabilitation system equipped with automated motion assessment depicted in Fig. 1. At home, a patient performs the task of grasping an object placed on a table, i.e., a reach-to-grasp task [9]. The patient uses the mobile application to record a video of performing the task and upload the video to the system. An automated model estimates a score for the motor function after analyzing the video data. The patient receives recommendations for self-managed exercises or receives a progress report related to the patient's score. The key component of this system is the automated evaluation of upper extremity function, and there are two main technical challenges: (1) video is the only available modality, i.e., other equipment such as depth or inertial measurement unit (IMU) sensor is unlikely to be available in the home; (2) patients perform *identical* tasks, but the model needs to identify subtle differences in task execution, which makes the problem more challenging than typical action recognition tasks.

In this paper, we consider the problem of designing self-rehabilitation systems and developing an accurate video-based assessment of motor skills. To that end, we propose a deep learning-based model for automated motion assessment and design a mobile application based on the algorithm. Our model evaluates the upper limb function of stroke survivors using only video data without additional sensors. The development of a fine-grained motion assessment model that performs accurate and reliable evaluations using only video data is the main goal of this research. We propose leveraging two modalities, RGB and optical flow, extracted from the video in order to capture subtle details in the execution of reach-to-grasp tasks at the pixel level. We developed a deep learning model that effectively learns the association between the two modalities by properly mixing features and subsequently applying cross-attention based on the Transformer architecture. We evaluated the performance of our model using a dataset we created and it contains 793 video clips of the reach-to-grasp motion of stroke survivors. Our experiments show that the proposed model achieves significantly higher accuracy than existing prior state-of-the-art (SOTA) methods for action recognition.

In addition, we developed a mobile application based on the proposed model for automated assessment. The dysfunction of upper extremity and its degree varies widely among patients. Taking the variability into account, we developed a total of 57 exercises in four category types: postural or balance exercise, range of motion exercise, strengthening exercise, and task-oriented training, with three levels of difficulty based on the severity of impairments. Our mobile application has several benefits. First, it automatically recommends exercises every day according to the motor function level of each patient. Second, experts in stroke rehabilitation participated in the application's development, and third, the application recommends exercises which the patient can easily and safely perform on their own in the sitting position. Moreover, a self-rehabilitation approach using mobile applications has many advantages, including the absence of professional guidance and supervision requirements, and the ability for patients to exercise without significant time and space constraints. Mobile applications also enable patients to assess their motor function and practice appropriate exercises. These self-rehabilitation approaches are less costly and allow patients to spend more time exercising.

The main contributions of our work are summarized as follows:

1) Our model assesses the motor function for stroke rehabilitation using only videos.
2) An accurate algorithm for motor function assessment was developed leveraging a novel method of modality fusion, which combines RGB and optical flow data based on deep learning.
3) Our model achieved higher accuracy in the assessment of upper limb functions over SOTA methods.
4) A mobile health application was developed, which recommends various types of self-administered exercises based on the assessment results and provides reports on their rehabilitation progress.

II. Related Work

A. Mobile Health Application for Stroke Rehabilitation

Smartphones are increasingly used by the general population, making it relatively easy to implement treatment programs at low costs. In the field of rehabilitation, various applications have been reported for each theme, such as stroke, traumatic brain injury, spinal cord injury, musculoskeletal, cardiac, pulmonary, cancer, and pain [10]. In stroke rehabilitation, mobile applications have been developed for specific goals, such as improvement of feedback for physical activity, aphasia training, cognitive assessment, training for patients with unilateral spatial neglect, education for home-based exercise, and functional skill training [11].

The effectiveness of mobile health programs was studied by Chung et al. [12] which showed that a mobile video-guided home exercise program for stroke patients has a higher self-efficacy and exercise adherence than paper-based programs. Several studies focused on increasing and promoting physical activity in patients. For example, an evidence-based behavior change technique was used through interactive mobile applications [13], and a finger training app on tablet PCs was developed to restore the ability to use the affected hands of stroke patients [14]. Ballard et al. [15] developed a language therapy application to improve the word-production ability of stroke patients suffering from apraxia of speech and aphasia.

Recently, there has been growing interest on mobile applications providing stroke rehabilitation programs on language and speech skills, physical therapy, and exercises [16], [17]. Some applications were developed for upper extremity rehabilitation. For example, studies in [18] and [19], developed software systems in which mobile applications are coupled with objects generated by 3D printers, resulting in high efficacy in home-based upper limb rehabilitation. Rehabilitation treatment programs in mobile game-based virtual reality were shown to be effective in promoting the recovery of upper limb function in stroke patients [20], [21], [22]. Most upper limb exercises can be performed while sitting down and, with a suitable guide, pose no significant safety risks even if performed on their own by the patients. Therefore, it is appropriate to develop an upper limb exercise program as a mobile health application. The unique features of our application, such as automated action evaluation and personalized exercise recommendation, are expected to be of great help to stroke patients.

B. Automated Assessment of Motor Function

The use of various sensing equipment in the automated assessment of motor function in neurorehabilitation has recently been explored [23], [24], [25], and a description of representative sensors is provided in Table I. Joint tracking data from Kinect [26] have been used to evaluate the motor function of patients using machine learning algorithms [27], [28]. Data from additional sensors such as IMU sensors [29] and force sensing resistors [30] were integrated with Kinect data to analyze the patients' movements. sEMG signals collected during daily activities have been used to evaluate patients' Brunnstrom stage of recovery [31]. A home rehabilitation system [32] has been developed using a smartwatch to collect IMU accelerometer and gyroscope data. In addition, several works studied the deep learning-based action recognition for post-stroke rehabilitation using IMU sensors [33], [34] or Kinect [35], [36]. However, the aforementioned studies used sensor equipment, which limits their widespread usage in home rehabilitation. Unlike these studies, our model only requires a (smartphone) camera, helping patients perform self-administered rehabilitation without significant limitations in equipment.

C. Deep Learning for Action Recognition

The task of action recognition in videos has been extensively studied using deep learning. I3D [37] has proposed to inflate 2D convolutional filters and pooling into 3D filters, and it uses two streams of data: RGB and optical flow data. ResNet3D [38] has extended the popular ResNet architecture to 3D data to prevent overfitting. ResNeXt3D [39] has applied a grouped convolution to the bottleneck module of ResNet3D to improve efficiency of the model. X3D [40] has explored the accuracy-complexity trade-off by expanding 2D data at various temporal and spatial dimensional scales. TDN [41] has captured local and global temporal information by learning short and long-term differences in motion. TimeSformer [42] has proposed a Divided Space-Time Attention which applies self-attention to the temporal feature of the video, as well as the spatial feature. MViT [43] has applied a 4-stage scale hierarchy to Vision Transformer where the deeper the stage, the lower the spatial resolution of the feature and the higher the channel dimension. In addition, there have been studies on fine-grained action recognition for automated evaluation [44], [45], [46] that focused mostly on discriminating between different types of action tasks with a wide range of motion. In contrast, we address the problem of assessing identical tasks which show relatively small differences across participants. This is achieved by proposing a novel modality fusion method that combines RGB and optical flow data to detect small changes in motion at the pixel level.

TABLE I
DESCRIPTION OF VARIETY OF SENSORS FOR REHABILITATION PURPOSE

Sensors	Description	Rehabilitation purpose
Kinect	Kinect is a motion-sensing device developed by Microsoft that tracks a user's movements in three-dimensional space. It consists of an RGB camera, a depth sensor, and multiple array microphones. The depth sensor consists of an infrared projector and a mono CMOS sensor to enable 3D imaging.	In motor rehabilitation, Kinect is used to accurately track a patient's movements and analyze them in real-time to monitor the rehabilitation process. This allows physical therapists to quantitatively assess and improve a patient's gait, balance, and other motor functions.
IMU	IMU sensors combine accelerometers, gyroscopes, and magnetometers to measure a device's linear velocity, rotation, and gravitational orientation. They come in a variety of sizes and shapes and are used for gait analysis, posture tracking, etc.	IMU sensors are used to precisely measure and analyze the movement and posture of stroke patients. This allows for precise assessment of gait patterns, balance, and rotational movements, as well as monitoring the progress of rehabilitation training.
sEMG	sENG is a non-invasive measurement of the electrical signals generated by muscles through surface electrodes. The signals reflect a variety of conditions, including degeneration, damage, and fatigue of muscle fibers.	In stroke patients, sEMG is an important tool of rehabilitation; by quantitatively measuring muscle activity, physical therapists can assess a patient's muscle function and improve or adjust their training program.
FSR (Force Sensing Register)	FSR is a sensor that measures pressure or force. When a force is applied, the resistance changes. This change is converted into an electrical signal, which allows for a quantitative measurement of the magnitude of the pressure.	FSRs are particularly useful for analyzing gait in stroke patients. While traditional gait analysis methods rely primarily on visual assessments, FSR allows clinicians to quantitatively measure and analyze gait characteristics such as patients' foot contact patterns, weight shifts, etc.

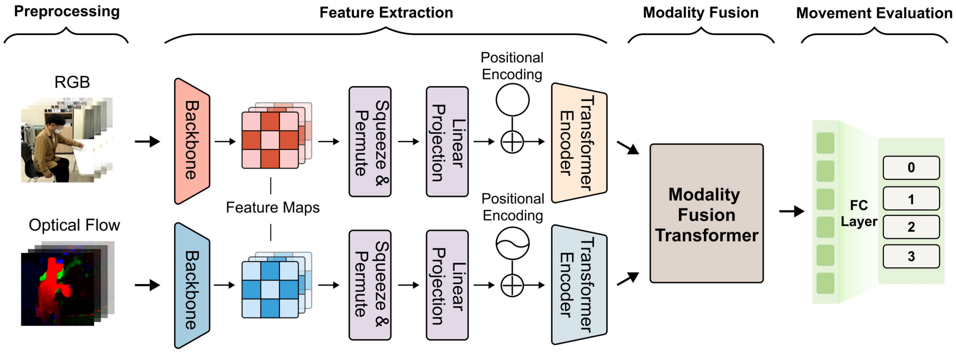


Fig. 2. Overall architecture of the proposed model.

III. Deep Learning Algorithm

A. Assessment of Upper Extremity Function

The videos of patients performing a reach-to-grasp task were recorded with the patients' consent. The task consisted of reaching for and grabbing a plastic cone that was placed on a table. This task was adopted from the Reaching Performance Scale (RPS) [9], which was developed to evaluate compensatory movements in the upper extremity during reaching and grasping tasks [9], [47], [48]. In RPS, the following six components are evaluated: *trunk displacement, movement smoothness, shoulder movements, elbow movements, prehension*, and *global score*. We used the *global score*, which evaluates the global quality of movements in the upper limb. According to the RPS, the global score has four levels [9]:

- Score 0: Less than half the task is accomplished despite modifications.
- Score 1: The task is done partially ($\geq$ 50%) or with modification (such as stabilization of the cone, sliding the cone on the table, modification of table height, shorter distance to the cone). Prehension may be absent.
- Score 2: The task is done in the presence of tremor; dysmetria; small, jerky movements; arc-shaped trajectory or segmentation. Prehension is possible but may be modified or difficult.
- Score 3: The task can be done easily, with or without mild tremor or dysmetria, following a smooth and direct trajectory.

B. Deep Learning Model for Automated Assessment

We present a deep learning model for classifying patients performing reach-to-grasp tasks into the four global score levels. The model uses only patients' videos as input, and it does not use any additional equipment such as depth or IMU sensors. Another challenge for the model is that the patients perform *identical* tasks. Thus, the differences between movements that are classified to different levels are subtle, which is not the case for typical action recognition tasks, such as differentiating between jumping and sleeping [37]. It would be beneficial to detect small changes in movement at the pixel level. Thus, in addition to the RGB data, we propose using *optical flow* data [49], i.e.the rate of change in pixel values in the 2D field, as complementary information to the RGB data.

As depicted in Fig. 2, our model consists of four stages: preprocessing, feature extraction, modality fusion, and classification.

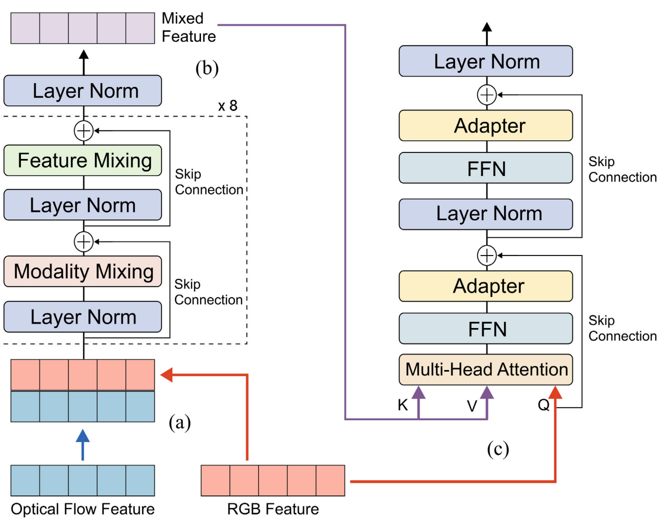
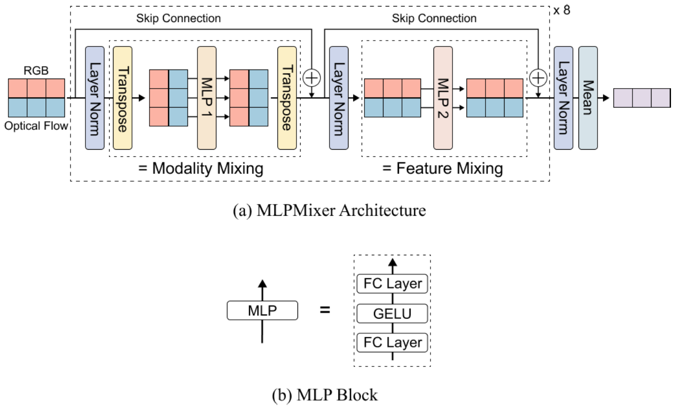


Fig. 3. Overall architecture of the proposed model combining MLPMixer and cross-attention for modality fusion. (a) The RGB and optical flow features are concatenated and fed into the MLPMixer. (b) The MLPMixer performs modality and feature mixing to generate the global mixed feature. (c) We used the mixed feature as Key, Value, and the RGB feature as Query of Cross-Attention module with the Adapter layer.

1) *Preprocessing:* The inputs to the model are the RGB and optical flow data, both of which are extracted from the video. The collected videos are of varying durations, depending on the patients' performance levels, and range from 36 to 441 frames in length. Patients who achieved high scores required a relatively short amount of time, whereas patients with low scores required more time because they felt relatively uncomfortable with the task. However, the frame length of the input videos had to be identical for the training of the model. Thus, a frame sampling was performed similarly to the process in a previous study [50], and the input frame length was fixed to 100 after preprocessing.

2) *Feature Extraction:* To obtain high-level spatio-temporal feature from the input, we first extracted a feature map using the ResNet3D-50 backbone [38], which is widely used in action recognition. A ResNet3D model was pre-trained on the Kinetics dataset [37], which contains 650,000 video clips of human actions. Subsequently, a Transformer encoder [51] was used to capture the latent semantic and global dependency of the spatio-temporal feature output by the backbone.

The output feature from the backbone undergoes several transformations before being input to the Transformer encoder as follows. A feature map output by the backbone has dimensions $C \times T \times H \times W$, which represents T frames of size $C \times H \times W$, where C, H, W, and T denote channel, height, width, and temporal dimension, respectively. The spatial dimension is squeezed out by the 2D spatial average pooling to change the feature dimensions to $C \times T$ which is permuted to the dimensions $T \times C$. Each feature vector in the time dimension is then projected to an h-dimensional vector. As a result, the output feature has the dimensions $T \times h$, which represents T tokens of feature of size h. The token sequence is input into the Transformer encoder after positional encoding is applied to it. A basic Transformer block [51] was used for the Transformer encoder.



(a) MLPMixer Architecture

(b) MLP Block

Fig. 4. Overall architecture of MLPMixer [52]. (a) MLPMixer performs modality mixing and feature mixing. In the modality mixing step, the input is transposed to mix the tokens at the same position in the different modalities. The output of modality mixing is mixed in feature dimension in the feature mixing step. (b) In both mixing processes, the feature is processed by MLP blocks. The MLP block consists of two fully-connected layers and GELU.

3) *Modality Fusion:* We propose a novel module to effectively fuse the features extracted from the RGB and optical flow data. There exists an asymmetry in the modalities: RGB data is the main modality, whereas optical flow is a modality derived from RGB data. Our model is designed to capture such asymmetry, and attempt to extract information mainly from RGB feature; it uses optical flow feature as *contextual* information.

As shown in Fig. 3, our model fuses RGB and optical flow data in two steps. First, the features of two modalities output from Transformer encoders are concatenated and passed to the MLPMixer [52]. The MLPMixer first mixes the concatenated feature across modality dimensions (modality mixing) to generate the intermediate feature and then further mixes the intermediate feature (feature mixing). The details for the architecture of the MLPMixer are provided in Fig. 4.

Second, the mixed feature is passed as key-value pair to the cross-attention module [51], and the RGB feature is passed as a query. The assignment of query, key, and value is consistent with our design intention such that RGB is the main modality, whereas the mixed modality that contains optical flow is used for contextual feature [51]. Thus, our model learns the association between the modalities through cross-attention. The adapter layer [53] is then added for parameter-efficient tuning of the Transformer block. Finally, classification is performed on the output from the modality fusion module using the fully connected layer.

4) *Loss Function:* We applied PolyLoss [54], which is based on cross-entropy as the loss function. The cross-entropy quantifies the discrepancy between the output confidence of the model and the ground truth of the data. PolyLoss generalizes the cross-entropy by expanding the loss function with Taylor polynomials. One recommended type of PolyLoss is the first-order Taylor polynomial combined with cross-entropy, which is called Poly-1 loss [54]. The cross-entropy loss and Poly-1 loss are given by

$$L_{\mathrm{CE_i}} = -\log(p_i) \quad (1)$$

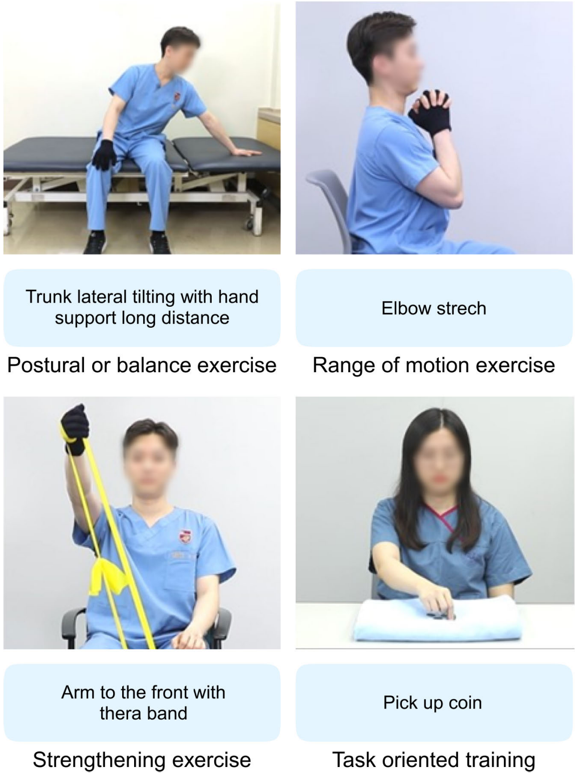

Trunk lateral tilting with hand support long distance
Postural or balance exercise
Elbow strech
Range of motion exercise
Arm to the front with thera band
Strengthening exercise
Pick up coin
Task oriented training

Fig. 5. Sample exercises in four categories.

$$L_{\mathrm{POLY}} = \sum_{i=1}^{N} \{ L_{\mathrm{CE_i}} + \epsilon_1 (1 - p_i) \} \tag{2}$$

where N denotes the number of video samples, p_i denotes the prediction confidence for the target ground truth score given input video i, and $L_{\mathrm{CE_i}}$ is the cross-entropy loss for input i. The hyperparameter ϵ_1 is called the perturbation coefficient. Poly-1 loss has shown better performance empirically than the cross-entropy or focal loss on several benchmarks for image classification [54].

IV. Development of Mobile Application

A mobile application was developed to support self-administered upper extremity exercise based on the automated evaluation by the proposed deep learning model. The degree of upper limb function after stroke varies widely, from the extent that it is impossible to perform any functional movements, to the extent that it is somewhat inconvenient but functional movements can be performed. Therefore, it is necessary to develop an exercise program based on each patient's upper limb disability after proper evaluation. Fig. 1 shows an overview of how users are evaluated for their own exercise level using the automated model proposed in the previous section and receive the exercise recommendations. In our application, we used a four-stage exercise level based on reach-to-grasp task.

A. Features of The Mobile Application

The mobile application contains various features supporting: my information, evaluation section, daily exercise, list of exercises, my current exercise status, and notification. The features are explained in detail with screenshots: see Fig. 6 and its caption.

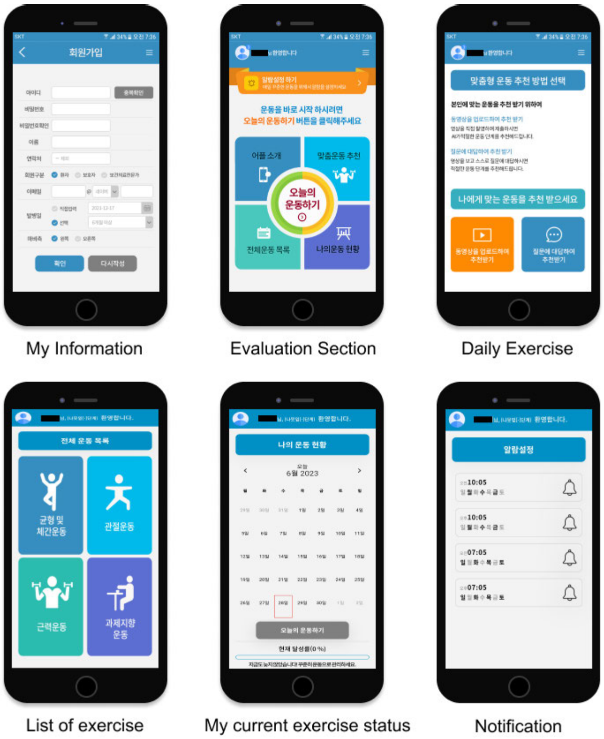


Fig. 6. Screenshots of proposed mobile application. (a) My information: This menu includes the user's personal details, such as user category (patient, caregiver, health professional), date of onset, and paralyzed side. (b) Evaluation section: This menu evaluates the user's upper extremity function using two methods: uploading a video of reach-to-grasp or answering a questionnaire. (c) Daily exercise: This menu shows five daily exercises each day. (d) List of exercises: This menu shows a list of exercises in four categories. (e) My current exercise status: This menu shows the exercise status in a calendar format. (f) Notification: The user can set a notification alarm on this page for routine exercises.

B. Types of Exercises for Upper Extremity

A total of 57 exercises in four category types (postural or balance exercise, range of motion exercise, strengthening exercise, and task-oriented training) were developed. The exercise program was divided into four exercise levels to accommodate different levels of upper extremity function and was designed as a 1-month program for patients.

The recommended exercises consisted of postural or balance exercises (n = 9), range of motion exercises (n = 7), strengthening exercises (n = 16), and task-oriented trainings (n = 25). Postural or balance exercise involves trunk flexion and weight-shifting. Range of motion exercise is performed on the shoulder, elbow, wrist, and finger joints. Strengthening exercise requires sandbag, dumbbell, thera band, handgrip equipment, socks, and hair tie for resistance. Task-oriented training is completed using cups, towels, and buttoned shirts, which are commonly found in everyday life. The exercises consist of bimanual activity, one-hand activity, and in-hand manipulation. A list of all the recommended exercises is provided in Table II. A group of specialists consisting of physical

TABLE II
LIST OF 57 RECOMMENDED EXERCISES IN FOUR CATEGORIES

Categories	Difficulty Levels	Exercises
Postural or Balance Exercise	1	Two arms push up, One arm push up, Trunk rotation, Abdominal drawing, Pelvic anterior and posterior tilting
	2	Trunk lateral tilting with hand support_short distance
	3	Trunk lateral tilting with hand support_long distance, Trunk lateral tilting with elbow support, Arm reaching in three direction
Range of Motion Exercise	1	Arm stretch to the front, Arm stretch to the side, Shoulder rotation, Elbow strech, Forearm supination/pronation, Wrist stretch, Finger stretch
Strengthening Exercise	1	Arm to the front with bean bag, Arm to the side with bean bag, Elbow exercise with bean bag, Wrist exercise with bean bag, Hand squeeze
	2	Arm to the front with dumbbell, Arm to the side with dumbbell, Arm to the front with thera band, Arm to the side with thera band, Arm rotation with thera band, Elbow exercise with dumbbell, Elbow exercise with thera band, Wrist exercise with dumbbell, Hand exercise with gripper
	3	Arm to the behind with thera band, Finger stretch with hair tie
Task Oriented Training	1	Sliding using two hands, Moving using two hands, Cup moving using two hands, Folding towel using strong hand, Draw the tissue using strong hand, Buttoning using strong hand, Moving bean bag
	2	Folding towel using weaker hand, Draw the tissue using weaker hand, Take off the lid using strong hand, Beading using strong hand, Cup moving, Pick up ball, Turn the bookshelf
	3	Button using weaker hand, Take off the lid using weaker hand, Beading using weaker hand, Pick up coin, Put the coin down, Put the ball down, Coin flip, Card flip, Turn the lid, Writing, Use chopsticks

therapists, occupational therapists, and rehabilitation medicine doctors with more than 10 years of experience brainstormed to determine exercises which patients with hemiplegia could perform safely without help. The specialists selected exercises which can be safely performed while sitting based on the motor function score. Exercise for motor function Score 0 primarily includes passive joint exercises and trunk exercises, as most patients at this stage are unable to extend their arms on their own. Score 1 exercises are for patients who can extend their arms but have difficulty using their hands properly. Thus, the exercises consist of active shoulder and elbow joint exercises, passive wrist and finger joint exercises, and trunk exercises. Score 2 corresponds to patients whose arms can be extended and hands can be used but there are challenges in fine movements. Thus, task-oriented training which emphasizes fine motor skills in addition to active joint and trunk exercises is included. Exercises for Score 3 are similar to those in Score 2, but are somewhat more difficult and include strength training. Fig. 5 shows samples of the recommended exercises by each category. Our application recommends five exercises on a daily basis. The difficulty levels are chosen based on the assessed scores of motor function as follows. For patients with score 0, only the exercises with difficulty level 1 is recommended. For patients with score 1, exercises of difficulty levels 1 and 2 are recommended in the ratio 3:2. For patients with score 2, exercises with difficulty levels 1, 2, and 3 are chosen in the ratio 2:2:1. For patients with score 3, exercises with difficulty levels 1, 2, and 3 are recommended in the ratio 1:2:2.

C. Security and Privacy Considerations

For data security and privacy, users' authorization to access the application was controlled on the login page, which appears first on opening the mobile application. To gain access, the page requires the users to enter an assigned unique email address and password. The data from the mobile application is transferred from the application to cloud storage through encryption, which makes the data illegible and unusable to unauthorized persons.

TABLE III
GENERAL CHARACTERISTICS AND CLINICAL DATA OF PARTICIPANTS

Variables	
Sex (n)	68 male, 32 female
Age	63.95 ± 13.22 years (28 ∼ 85)
BMI	23.14 ± 3.19 (14.69 ∼ 39.06)
Type of stroke	
Ischemic	72
Hemorrhagic	28
Side of hemiplegia (n)	right 52, left 48
Time after post stroke (weeks)	100.54 ± 193.31 (1.14 ∼ 1273.86)
NIHSS	5.49 ± 3.03 (1 ∼ 14)
FMA for upper limb	47.74 ± 19.69

BMI, body mass index; NIHSS, National Institute of health stroke scale; FMA, Fugl-Meyer Motor Assessment

V. EXPERIMENT

A. Setup

1) Dataset: A total of 100 stroke survivors with hemiplegia were recruited from Korea University Guro Hospital and Sahmyook Medical Center. The participants needed to meet the following inclusion criteria: have hemiplegia due to ischemic or hemorrhagic stroke, be aged $>$ 18 years, have given their informed consent, and have adequate cognitive function to understand the instructions and perform the tasks appropriately. The NIHSS and Fugl-Meyer Motor Assessment (FMA) [6], [25], [27], [29], [30], [32] of the upper extremity were used. This protocol was approved by the regional Institutional Review Board of the Korea University Guro Hospital (IRB No. 2021GR0178).

We obtained 793 videos of 100 stroke patients performing reaching and grasping tasks using a smartphone (iPhone 13 Pro). Most patients performed a reaching task for a far and a close target from four predetermined angles. Thus, approximately eight videos were obtained per patient. Two specialists (an occupational therapist and a physiatrist) independently evaluated the videos, and each patient's exercise score was

TABLE IV
COMPARISON BETWEEN OUR MODEL AND THE SOTA MODELS IN ACTION RECOGNITION. THE RESULTS WERE AVERAGED OVER 20 REPETITIONS OF EXPERIMENTS

Method	Modality	Accuracy (%)
I3D [37]	RGB, Optical Flow	64.50 ± 3.38
ResNet3D [38]	RGB	78.50 ± 1.89
ResNeXt3D [39]	RGB	78.25 ± 3.52
X3D [40]	RGB	75.42 ± 2.96
TDN [41]	RGB	82.58 ± 1.56
TimeSformer [42]	RGB	60.33 ± 2.08
MViT [43]	RGB	79.92 ± 2.87
Proposed	**RGB, Optical Flow**	**86.08 ± 1.97**

TABLE V
(A) CONFUSION MATRIX ASSOCIATED WITH THE PREDICTED SCORE (B) AVERAGE FMA SCORES ASSOCIATED WITH THE PERFORMANCE SCORE OF PATIENTS

(a)

		Predicted Score			
		0	1	2	3
True Score	0	**14**	2	0	0
	1	0	**13**	0	3
	2	0	1	**26**	5
	3	0	0	0	**56**

(b)

Motor Function Score	FMA score
0	13.44 ± 10.43
1	31.27 ± 11.74
2	49.21 ± 17.52
3	61.08 ± 6.97

TABLE VI
ABLATION STUDY

Method	Accuracy (%)
Optical flow-only	70.58 ± 1.83
RGB-only	84.83 ± 1.97
Modality-fusion Transformer [55]	84.42 ± 2.44
Proposed	**86.08 ± 1.97**

classified into one of four levels according to the global score criterion in the RPS [9]. The intraclass correlation coefficient of two specialists was high (0.972). If the two specialists did not agree, another specialist evaluated that video and assessed the patients score. After a discussion of three specialists, the patient's score was determined. A total of 673 videos were used for training, and 120 videos were used for the performance evaluations. Details of the participants are provided in Table III.

2) Implementation Details: For both modalities of RGB and optical flow, the channel dimension of the output from ResNet3D-50 backbone is 2048. The input of the backbone model has a shape of 3 × 100 × 256 × 256. The output is projected to 768-dimensional space using one FC (fully connected) layer. For the Transformer encoder, the number of encoder layers is 3, the number of attention heads is 12 and the feedforward dimension is set as input dimension × 4. The cross-attention module has 12 attention heads, and the dimension of hidden states is 3072. The dropout rates of positional embedding, Transformer encoders, and cross-attention are all set to 0.2. The classification network is constructed with an FC layer with an input dimension of 768. Our model was trained on one NVIDIA RTX 3090 GPU for 200 epochs with a batch size of 2. We used SGD(stochastic gradient descent) optimizer, and the initial learning rate was set to 0.001.

B. Classification Performance

1) Baseline Methods: To demonstrate the effectiveness of the proposed model in discriminating between fine-grained movements by patients, we compared our model with four existing SOTA action recognition models and video classification models as the baselines: I3D [37], ResNet3D [38], ResNeXt3D [39], X3D [40], TDN [41], TimeSformer [42], and MViT [43].

2) Results and Discussion: As shown in Table IV, the overall accuracy of the proposed method was 86.08%. The Pearson correlation coefficient between the actual and the predicted scores showed a high correlation with $R = 0.93$, $p < 0.001$. Compared with ResNet3D, ResNeXt3D, X3D, TDN, TimeSformer, and MViT which used only RGB data, the proposed model showed a significant improvement in accuracy using the combined RGB and optical flow data. Notably, I3D also used a two-stream network that combined RGB and optical flow data [37]. However, our model outperformed I3D by a large margin. This result demonstrates that simply combining two streams is insufficient, and the proper fusion of the streams is crucial for the fine-grained evaluation of motor function.

Table V(a) presents the confusion matrix associated with the prediction output of our model. The degree of confusion was relatively high between the two highest scores, i.e., score values of 2 and 3. The patients who achieved high scores were able to accomplish the task in a relatively short time. Thus, the variability in motion across those videos would be low, which makes them more difficult to classify. By contrast, the motion of patients with low scores would show greater variability, because they tend to perform the task with hesitation or difficulty [9]. This observation is supported by the results in Table V(b) which shows the average FMA scores associated with the performance scores of participants. The construct validity between RPS and FMA is strong (Spearman rho = 0.88-0.89, p < 0.0001) [47]. In the table, the gap between FMA scores for participants of scores 0-1 and 1-2 is approximately 18, however, the gap in the FMA score decreases to about 12 for participants with scores 2-3. Thus, the difference in motor skills for participants with high scores is potentially small. In addition, our model rarely made a prediction that was incorrect by 2 or more score points, i.e., there were only 3 such cases in total of 120 videos.

3) Ablation Study: We performed an ablation study on the following components in our model: (i) the use of the two modalities of RGB and optical flow data; (ii) the modality fusion network. We considered the following three baselines: (1) RGB-only; only the branch for feature extraction from RGB in Fig. 2 is used. (2) Optical flow-only; only the branch for feature extraction from optical flow in Fig. 2 is used. (3) Recently proposed model for modality fusion [55]; the RGB and optical flow features are fused using the multimodal fusion Transformer proposed in [55]. Table VI shows the performance accuracies for the ablation study. The results show that the multimodality approach and the proposed method for fusing multimodal features are the most effective for fine-grained motion assessment.

TABLE VII
MOBILE APPLICATION RATING SCALE

	App Quality Rating				App
	Engagement	Functionality	Aesthetics	Information	Subjective
Patient	3.94 ± 0.54	4.07 ± 0.67	3.73 ± 0.73	3.34 ± 0.41	16.10 ± 2.33
Therapist	3.98 ± 0.50	4.27 ± 0.62	4.00 ± 0.59	3.58 ± 0.27	16.00 ± 2.71
Physiatrist	3.40 ± 0.51	3.73 ± 0.57	2.97 ± 0.73	2.77 ± 0.61	12.80 ± 2.57
Total	3.77 ± 0.57	4.03 ± 0.64	3.57 ± 0.80	3.23 ± 0.56	14.97 ± 2.90

Values are mean ± standard deviation

C. Usability and Quality of Mobile Application

The mobile implementation was on Galaxy Note 10 device based on Android 12 and Exynos 9825 chipset. The inference of evaluation scores is performed on NVIDIA RTX 2080 GPU at the server. The latency of the overall process between the initiation of uploading videos and the display of scores was 9.5 seconds on average.

The Mobile Application Rating Scale (MARS) [56] was used to assess the usability and quality of the health mobile application. MARS contains five broad categories of criteria, including four objective quality scales: engagement, functionality, aesthetics, and information quality, and one subjective quality. In total, 30 people (10 patients, 10 physiatrists, and 10 therapists) rated the application using the MARS, and the results are shown in Table VII. The mean score of application quality was 3.64 ± 0.55 (perfect score = 5), and the mean subjective application score was 14.97 ± 2.90 (perfect score = 20).

VI. CONCLUSION

In this paper, we proposed a deep learning model which automatically evaluates stroke survivors' upper extremity function based only on videos in which the patient performs reach-to-grasp task. We adopted a multimodality approach to discriminate between subtle movements of patients performing identical tasks. The proposed model with modality fusion was effective in fine-grained assessment, and it significantly outperformed existing SOTA models for action recognition. Based on the proposed model, we developed a mobile application that supports self-administered upper limb exercise for patients with hemiplegia. Based on the MARS assessment criteria, this application has garnered favorable evaluations in terms of quality from its prospective user base.

This study has some limitations. First, we focused on developing a mobile application and its quality and usability. The effectiveness of the application was not investigated, and further long-term studies are required to address this limitation. Second, our application does not record the patient's exercise performance or provide feedback on the appropriateness of exercise performance. Further development of algorithms is needed to resolve such one-sidedness of our application.

In the future, we plan to enhance the assessment model by utilizing additional modalities of task execution of patients. In this work, we chose the modality of RGB and optical flow (derived from RGB) for the widespread adoption of our mobile application. However, if additional sensors or measurements (e.g., depth sensors, joint estimation) become more widely available, we aim to exploit such modalities in our subsequent research for more precise evaluation. In addition, we envision an advanced rehabilitation system that combines visual assessment of motion with numerical analysis of joint and muscle movements. This research will aim at developing an interpretable model which provides logical and explainable feedback for self-administered rehabilitation.

REFERENCES

[1] C.-M. Chen, C.-C. Tsai, C.-Y. Chung, C.-L. Chen, K. P. Wu, and H.-C. Chen, "Potential predictors for health-related quality of life in stroke patients undergoing inpatient rehabilitation," *Health Quality Life Outcomes*, vol. 13, no. 1, pp. 1–10, Dec. 2015.

[2] J. G. Broeks, G. Lankhorst, K. Rumping, and A. Prevo, "The long-term outcome of arm function after stroke: Results of a follow-up study," *Disability Rehabil.*, vol. 21, no. 8, pp. 357–364, Jan. 1999.

[3] G. Kwakkel et al., "Motor rehabilitation after stroke: European stroke organisation (ESO) consensus-based definition and guiding framework," *Eur. Stroke J.*, vol. 8, no. 4, pp. 880–894, Dec. 2023.

[4] E. J. Schneider, N. A. Lannin, L. Ada, and J. Schmidt, "Increasing the amount of usual rehabilitation improves activity after stroke: A systematic review," *J. Physiotherapy*, vol. 62, no. 4, pp. 182–187, Oct. 2016.

[5] K. R. Lohse, C. E. Lang, and L. A. Boyd, "Is more better? Using metadata to explore dose–response relationships in stroke rehabilitation," *Stroke*, vol. 45, no. 7, pp. 2053–2058, Jul. 2014.

[6] D. E. Feldman, "A self-administered graded repetitive arm supplementary program (GRASP) improves arm function during inpatient stroke rehabilitation: A multi-site randomized controlled trial," *Yearbook Sports Med.*, vol. 2010, p. 254, Jan. 2010.

[7] L. F. Chin, K. S. Hayward, A. L. M. Chai, and S. G. Brauer, "A self-empowered upper limb repetitive engagement program to improve upper limb recovery early post-stroke: Phase II pilot randomized controlled trial," *Neurorehabil. Neural Repair*, vol. 35, no. 9, pp. 836–848, Sep. 2021.

[8] K. Westlake, R. Akinlosotu, J. Udo, A. G. Shipper, S. M. Waller, and J. Whitall, "Some home-based self-managed rehabilitation interventions can improve arm activity after stroke: A systematic review and narrative synthesis," *Frontiers Neurol.*, vol. 14, Feb. 2023, Art. no. 1035256.

[9] M. F. Levin, J. Desrosiers, D. Beauchemin, N. Bergeron, and A. Rochette, "Development and validation of a scale for rating motor compensations used for reaching in patients with hemiparesis: The reaching performance scale," *Phys. Therapy*, vol. 84, no. 1, pp. 8–22, Jan. 2004.

[10] R. Nussbaum, C. Kelly, E. Quinby, A. Mac, B. Parmanto, and B. E. Dicianno, "Systematic review of mobile health applications in rehabilitation," *Arch. Phys. Med. Rehabil.*, vol. 100, no. 1, pp. 115–127, Jan. 2019.

[11] X. Zhou, M. Du, and L. Zhou, "Use of mobile applications in post-stroke rehabilitation: A systematic review," *Topics Stroke Rehabil.*, vol. 25, no. 7, pp. 489–499, Oct. 2018.

[12] B. P. H. Chung et al., "Pilot study on comparisons between the effectiveness of mobile video-guided and paper-based home exercise programs on improving exercise adherence, self-efficacy for exercise and functional outcomes of patients with stroke with 3-month follow-up: A single-blind randomized controlled trial," *Hong Kong Physiotherapy J.*, vol. 40, no. 1, pp. 63–73, Jun. 2020.

[13] L. Paul et al., "Increasing physical activity in stroke survivors using starfish, an interactive mobile phone application: A pilot study," *Topics stroke Rehabil.*, vol. 23, no. 3, pp. 170–177, 2016.

[14] S. H. Jang and W. H. Jang, "The effect of a finger training application using a tablet PC in chronic hemiparetic stroke patients," *Somatosensory Motor Res.*, vol. 33, no. 2, pp. 124–129, Apr. 2016.

[15] K. J. Ballard, N. M. Etter, S. Shen, P. Monroe, and C. T. Tan, "Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia," *Amer. J. Speech-Lang. Pathol.*, vol. 28, no. 2S, pp. 818–834, Jul. 2019.

[16] P. Piran et al., "Medical mobile applications for stroke survivors and caregivers," *J. Stroke Cerebrovascular Diseases*, vol. 28, no. 11, Nov. 2019, Art. no. 104318.

[17] S. P. Burns et al., “MHealth intervention applications for adults living with the effects of stroke: A scoping review,” *Arch. Rehabil. Res. Clin. Transl.*, vol. 3, no. 1, Mar. 2021, Art. no. 100095.

[18] H. Zhang et al., “RehabPhone: A software-defined tool using 3D printing and smartphones for personalized home-based rehabilitation,” in *Proc. 18th Int. Conf. Mobile Syst., Appl., Services*, Jun. 2020, pp. 434–447.

[19] S. Bhattacharjya et al., “Harnessing smartphone technology and three dimensional printing to create a mobile rehabilitation system, mRehab: Assessment of usability and consistency in measurement,” *J. NeuroEng. Rehabil.*, vol. 16, no. 1, pp. 1–13, Dec. 2019.

[20] Y.-H. Choi, J. Ku, H. Lim, Y. H. Kim, and N.-J. Paik, “Mobile game-based virtual reality rehabilitation program for upper limb dysfunction after ischemic stroke,” *Restorative Neurol. Neurosci.*, vol. 34, no. 3, pp. 455–463, Jun. 2016.

[21] Y.-H. Choi and N.-J. Paik, “Mobile game-based virtual reality program for upper extremity stroke rehabilitation,” *J. Visualized Exp.*, no. 133, Mar. 2018, Art. no. e56241.

[22] N. LaPiana et al., “Acceptability of a mobile phone–based augmented reality game for rehabilitation of patients with upper limb deficits from stroke: Case study,” *JMIR Rehabil. Assistive Technol.*, vol. 7, no. 2, Sep. 2020, Art. no. e17822.

[23] J. Ambros-Antemate et al., “Accuracy of machine learning algorithms for the assessment of upper-limb motor impairments in patients with post-stroke hemiparesis: A systematic review and meta-analysis,” *Adv. Clin. Experim. Med.*, vol. 31, no. 12, pp. 1309–1318, 2022.

[24] A. Arac, “Machine learning for 3D kinematic analysis of movements in neurorehabilitation,” *Current Neurol. Neurosci. Rep.*, vol. 20, no. 8, pp. 1–6, Aug. 2020.

[25] E. D. Oña Simbaña, P. Sánchez-Herrera Baeza, A. Jardón Huete, and C. Balaguer, “Review of automated systems for upper limbs functional assessment in neurorehabilitation,” *IEEE Access*, vol. 7, pp. 32352–32367, 2019.

[26] Z. Zhang, “Microsoft Kinect sensor and its effect,” *IEEE Multimedia Mag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[27] W.-S. Kim, S. Cho, D. Baek, H. Bang, and N.-J. Paik, “Upper extremity functional evaluation by Fugl–Meyer assessment scoring using depth-sensing camera in hemiplegic stroke patients,” *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0158640.

[28] A. Scano, A. Chiavenna, M. Malosio, L. Molinari Tosatti, and F. Molteni, “Kinect v2 implementation and testing of the reaching performance scale for motor evaluation of patients with neurological impairment,” *Med. Eng. Phys.*, vol. 56, pp. 54–58, Jun. 2018.

[29] P. Otten, J. Kim, and S. Son, “A framework to automate assessment of upper-limb motor function impairment: A feasibility study,” *Sensors*, vol. 15, no. 8, pp. 20097–20114, Aug. 2015.

[30] S. Lee, Y.-S. Lee, and J. Kim, “Automated evaluation of upper-limb motor function impairment using Fugl–Meyer assessment,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 125–134, Jan. 2018.

[31] L. Meng et al., “Automatic upper-limb Brunnstrom recovery stage evaluation via daily activity monitoring,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2589–2599, 2022.

[32] S. H. Chae, Y. Kim, K.-S. Lee, and H.-S. Park, “Development and clinical evaluation of a web-based upper limb home rehabilitation system using a smartwatch and machine learning model for chronic stroke survivors: Prospective comparative study,” *JMIR mHealth uHealth*, vol. 8, no. 7, Jul. 2020, Art. no. e17216.

[33] V. B. Semwal, A. Gupta, and P. Lalwani, “An optimized hybrid deep learning model using ensemble learning approach for human walking activities recognition,” *J. Supercomput.*, vol. 77, no. 11, pp. 12256–12279, Nov. 2021.

[34] V. B. Semwal, N. Gaud, P. Lalwani, V. Bijalwan, and A. K. Alok, “Pattern identification of different human joints for different human walking styles using inertial measurement unit (IMU) sensor,” *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 1149–1169, Feb. 2022.

[35] V. Bijalwan, V. B. Semwal, G. Singh, and R. G. Crespo, “Heterogeneous computing model for post-injury walking pattern restoration and postural stability rehabilitation exercise recognition,” *Expert Syst.*, vol. 39, no. 6, Jul. 2022, Art. no. e12706.

[36] V. Bijalwan, V. B. Semwal, G. Singh, and T. K. Mandal, “HDL-PSR: Modelling spatio-temporal features using hybrid deep learning approach for post-stroke rehabilitation,” *Neural Process. Lett.*, vol. 55, no. 1, pp. 279–298, Feb. 2023.

[37] J. Carreira and A. Zisserman, “Quo Vadis, action recognition? A new model and the kinetics dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.

[38] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3D residual networks for action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.

[39] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[40] C. Feichtenhofer, “X3D: Expanding architectures for efficient video recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 203–213.

[41] L. Wang, Z. Tong, B. Ji, and G. Wu, “TDN: Temporal difference networks for efficient action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1895–1904.

[42] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?” in *Proc. ICML*, vol. 2, no. 3, 2021, p. 4.

[43] Y. Li et al., “MViTv2: Improved multiscale vision transformers for classification and detection,” 2021, *arXiv:2112.01526*.

[44] T. Han, H. Yao, W. Xie, X. Sun, S. Zhao, and J. Yu, “TVENet: Temporal variance embedding network for fine-grained action representation,” *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107267.

[45] J. Hong, M. Fisher, M. Gharbi, and K. Fatahalian, “Video pose distillation for few-shot, fine-grained sports action recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9254–9263.

[46] A. Piergiovanni and M. S. Ryoo, “Fine-grained activity recognition in baseball videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1740–1748.

[47] F. P. P. V. de Andrade, R. S. Padula, A. C. Binda, M. L. da Silva, and S. R. Alouche, “Measurement properties of the reaching performance scale for stroke,” *Disability Rehabil.*, vol. 43, no. 8, pp. 1171–1175, Apr. 2021.

[48] S. K. Subramanian, M. C. Baniña, A. Turolla, and M. F. Levin, “Reaching performance scale for stroke—Test-retest reliability, measurement error, concurrent and discriminant validity,” *PM&R*, vol. 14, no. 3, pp. 337–347, Mar. 2022.

[49] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Performance of optical flow techniques,” *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, Feb. 1994.

[50] L. Wang et al., “Temporal segment networks for action recognition in videos,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[51] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[52] I. Tolstikhin et al., “MLP-mixer: An all-MLP architecture for vision,” in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[53] N. Houlsby et al., “Parameter-efficient transfer learning for NLP,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[54] Z. Leng et al., “PolyLoss: A polynomial expansion perspective of classification loss functions,” 2022, *arXiv:2204.12511*.

[55] N. Shvetsova et al., “Everything at once—Multi-modal fusion transformer for video retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19988–19997.

[56] S. R. Stoyanov, L. Hides, D. J. Kavanagh, O. Zelenko, D. Tjondronegoro, and M. Mani, “Mobile app rating scale: A new tool for assessing the quality of health mobile apps,” *JMIR mHealth uHealth*, vol. 3, no. 1, p. e27, Mar. 2015.