Prediction of upper limb function from simple activity of daily living using deep learning in patients with stroke

Dain Shim

Department of Medicine

The Graduate School, Yonsei University

Prediction of upper limb function from simple activity of daily living using deep learning in patients with stroke

Directed by Professor Dong-wook Rha

Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science/

Dain Shim

December 2023

This certifies that the Doctoral Dissertation of
Dain Shim is approved.

Thesis Supervisor : Dong-wook Rha

Thesis Committee Member#1 : Seo Yeon Yoon

Thesis Committee Member#2 : Ja Young Choi

Thesis Committee Member#3: Seong-A Lee

Thesis Committee Member#4: Young Dae Kim

The Graduate School
Yonsei University

December 2023

ACKNOWLEDGEMENTS

First, I would like to express my deep respect and gratitude to my supervisor, Prof. Dong-wook Rha, who taught me to take a broader view of my research and guides me in my research concerns.

And I would also like to thank my other thesis committee members, Prof. Seo Yeon Yoon, Prof. Ja Young Choi, Prof. Seong-A Lee, and Prof. Young Dae Kim who advised me about how to think broadly about convergence research asked such challenging questions to expand the way I think. Their kind and warm encouragement gave me great source of strength in conducting the research.

I would also like to thank my colleagues at the Rehabilitation Medicine Research Institute who supported me throughout my graduate school life.

Lastly, I am so grateful to my family who always love and support me. I will live my whole life giving back to my family who always worried about me, prayed for me, and rejoiced me in times of hardship and joy.

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Prediction of upper limb function from simple activity of daily living using deep learning in patients with stroke

Dain Shim

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Dong-wook Rha)

Upper limb function in stroke patients is commonly determined by clinical measurements such as the Fugl-Meyer Assessment-Upper Extremity (FMA-UE) and Box and Block Test (BBT), which are time-consuming and require trained clinicians. Three-dimensional (3D) computerized motion analysis is one alternative, but it is also time-consuming and requires expensive devices. So, we wanted to know if we could predict upper limb function from simple activity of daily life using deep learning. The aim of this study was to predict upper limb function in stroke patients using deep learning with short two-dimensional (2D) videos of patients performing a simple activity of daily living. To achieve this, we first developed models to predict metrics representing upper limb function using 3D motion capture data of patients with stroke. We then developed similar models to predict the same metrics using keypoints in 2D video data of patients with stroke.

We collected FMA-UE score, BBT score, the temporospatial parameters including Movement Time (MT), Index of Curvature (IC) and Number of Movement Units (NMU) and Arm Profile Score (APS) from 3D motion capture in 265 stroke patients from 2014 to 2023. In addition, 2D video data recorded during Reach & Grasp Cycle were collected in 103 stroke patients from 2021 to 2023. Two versions of input data were used to train the deep learning model. First, we used 3D coordinate data to construct the 3D motion capture dataset to predict metrics representing upper limb function. During 3D motion capture, we obtained a total of 30 coordinate data per trial, consisting of X, Y, and Z coordinates data of 10 reflex markers: Trunk (4), Shoulder, Elbow, Wrist (2), Finger (2). Second, we used 330 video clips to construct a 2D video dataset to predict metrics representing upper limb function. 2D keypoints were extracted through pose estimation using the RTMPose method. We obtained a total of 14-coordinate data per video, consisting of X and Y coordinates of 7 keypoints of upper limb from 2D video; Trunk, Shoulder, Elbow, Wrist (2), Finger (2). The Convolutional Neural Network (CNN) and Temporal Convolutional Network (TCN) were used to classify FMA and BBT into 3 groups by severity of upper limb dysfunction and to estimate temporospatial parameters and APS. The input data were divided into a training set (60%), a validation set (20%), and a test set (20%).

We found that a CNN performed better than a TCN for all predictions regardless of whether 3D or 2D data were used. The CNN model using 3D data had accuracy, precision, recall, and F1-score exceeding 90 for FMA-UE (91.13, 90.27, 90.35 and 90.31, respectively) and 72 for BBT prediction (79.03, 72.54, 73.96 and 73.24, respectively). The predicted MT, IC, NMU and APS had moderate to strong correlations with true value (r=0.544, 0.755, 0.601 and 0.783). The performance metrics were similar, each exceeding 80 for FMA-UE prediction (89.23, 88.39, 85.97 and 87.16, respectively) and 73 for BBT prediction (76.92, 73.79, 75.51 and 74.64, respectively) when a CNN model was used with 2D data. The predicted MT, IC, NMU and APS had moderate to strong correlations with true value (r=0.528, 0.703, 0.625 and 0.569, respectively).

The deep learning method gave highly promising results in predicting upper limb function of stroke patients using only single 2D video recorded during simple activity of daily living. The upper limb dysfunction could be classified according to its severity according to FMA and BBT. Also, temporospatial parameters and APS showed moderate to strong correlation with the predicted values and true values.

Key words: stroke, upper limb function, prediction, deep learning, activity of daily living

Prediction of upper limb function from simple activity of daily living using deep learning in patients with stroke

Dain Shim

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Dong-wook Rha)

I. INTRODUCTION

Stroke is a major health problem and a leading cause of adult disability worldwide.[1] A stroke is caused by a burst or blockage of a blood vessel in the brain, which can result in loss or limitation of upper limb function.[2] Accurate measurement of upper limb function is important for confirming stroke patients' functional state, planning appropriate treatment, and improving motor function and quality of daily life.[3] Existing methods to evaluate upper limb function in stroke patients can be divided into two major categories. The first category consists of clinical methods in which clinicians or therapists observe a patient's movements and score upper limb function using evaluation tools whose reliability and validity have been proven. Examples of such tools include the Fugl-Meyer Assessment-Upper Extremity (FMA-UE), Box and Block test (BBT), and Jepsen Taylor hand function test.[4-7] However, these evaluations involve subjective judgments made by humans, so well-trained clinicians are needed to ensure their accuracy, and their results are semi-quantitative. The second major way to evaluate upper limb function in patients with stroke is by a three-dimensional (3D) computerized motion analysis test.[8] In this method, two or more infrared cameras record the movements of reflective markers attached to points on the subject's body, and the 3D coordinates of the markers are inversely calculated by triangulation after the subject is projected from the same point.[9] Although this method provides relatively objective and quantitative data, it requires expensive hardware and analysis software as well as a large space to install the hardware. In addition, highly skilled experts are needed to post-process the vast amount of data obtained and interpret the results. Since a vast amount of data is difficult to interpret easily, clinically meaningful parameters can be calculated using 3D motion capture data. In a previous study, Arm Profile Score (APS), which is a kinematic parameter, and spatiotemporal parameters including movement time, index of curvature, and number of movement units obtained from 3D motion analysis were found to have high correlations with upper limb function in children with cerebral palsy. However, to obtain these parameters, post-processing and a separate calculation process were required after the 3D motion capture.[10-11]

To overcome the limitations of existing methods for evaluation of upper limb function in patients with stroke, a new method using Artificial Intelligence (AI) technology is needed. AI and big data are currently being applied to many economic and social fields, resulting in innovative changes.[12] In particular, the development of AI is having a great impact on the medical field.[13-15] For example, AI is being used to interpret CT and MRI images and reduce doctors' reading times in Radiology Departments,[16] and to analyze physical function by using motion data for diagnosis or monitoring for functional recovery in Departments of Rehabilitation Medicine.[17] In addition, with the recent worldwide Covid-19 pandemic, interest in non-face-to-face or remote medical treatment has increased,[18-19] and several researchers have focused on using AI to improve the digital healthcare workflow.[20] Ongoing advances in AI have the potential to bring many more changes to medical diagnosis and treatment systems in the future.[21] These innovations are expected to transform traditional medical practices by increasing non-face-to-face patient diagnosis and treatment and providing personalized healthcare services effortlessly.[22]

The motion of objects can be detected and recognized by combining AI technology with image processing and analysis.[23] Furthermore, studies are being reported that estimate human motion and predict body functions using markerless motion capture from two-dimensional (2D) images.[24] For example, one study reported that gait metrics were well predicted using AI with 2D video recorded by a single camera.[25] Compared with marker-based motion capture, which is highly dependent on specialized hardware, markerless motion capture is inexpensive, independent of location, and easy to interpret because it does not require a complicated inspection process. We want to develop a convenient method that uses AI to measure body function based on 2D video of a subject's movement without trained experts and expensive equipment.

The ultimate goal of this study was to develop deep learning algorithms that predict upper limb function using 2D video data recorded while patients with stroke conduct only simple activities of daily living. To achieve this goal, three processes were performed. First, we developed deep learning algorithms to predict upper limb function using 3D motion capture data and explored the possibility of similarly predicting upper limb function with 2D video data. Second, to implement markerless motion capture, we used a pose estimation algorithm to accurately detect keypoints from 2D video. Third, we developed deep learning algorithms to predict upper limb function using 2D keypoint data estimated from 2D video.

II. MATERIALS AND METHODS

1. Participants

Participants were stroke patients visited to the Department of Rehabilitation Medicine in Severance rehabilitation hospital between October 2014 and September 2023 who underwent upper limb motion analysis test and clinical upper limb function evaluation. Inclusion criteria for the study were: (1) adults with stroke 18 years of age or older, (2) hemiplegic or quadriplegic patients, and (3) clinical assessment and 3D motion analysis performed within seven days. Patients were excluded if they met any of the following criteria: (1) had other musculoskeletal or nervous system disorders, (2) had insufficient cognitive function to follow the instructions for clinical assessment and 3D motion analysis, or (3) were judged by the researcher to be unsuitable for participation.

2. Study design

This retrospective study used clinical measurements and 3D motion capture data from 265 patients with stroke who underwent clinical evaluation of upper limb function and 3D motion capture-based analysis of upper limb motion in the Department of Rehabilitation Medicine at Severance Rehabilitation Hospital between October 2014 and September 2023. Clinical evaluation and 3D motion analysis of the upper extremities are performed as standard treatments for patients with stroke with upper limb functional impairment at Severance Rehabilitation Hospital. In addition, 2D video data were recorded simultaneously with 3D motion capture from October 2021 to September 2023. Ethical approval for this study was granted by the institutional review board and ethics committee (4-2023-0450). In our study, we used upper limb movement data to estimate parameters representing upper limb function that are currently being used in the hospital based on AI. Figure 1 summarizes the overall workflow of the study.

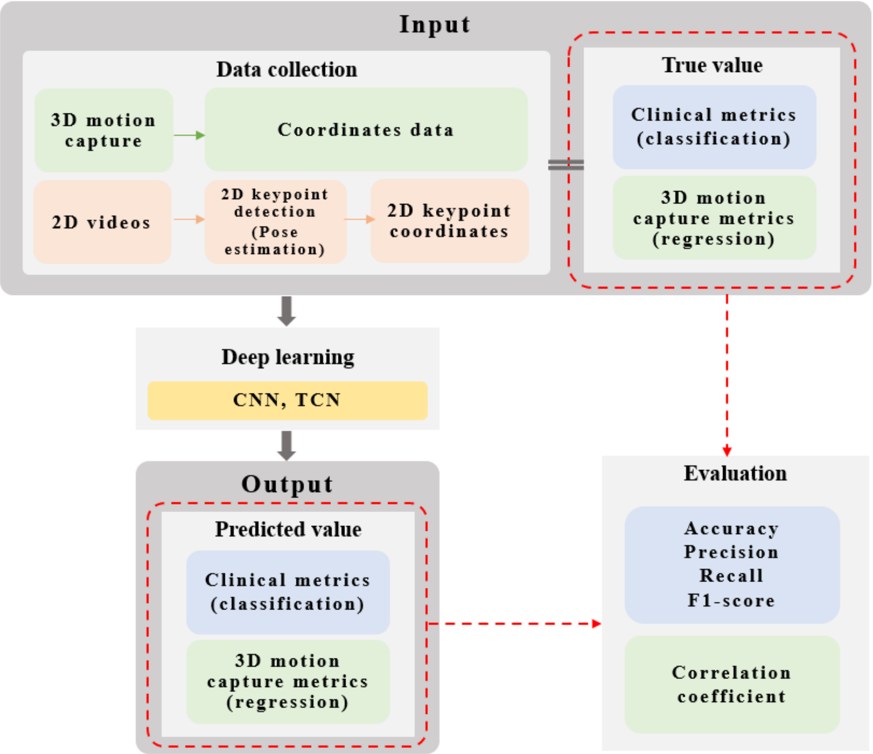


Figure 1. Overall workflow diagram

3. Data collection

A. Clinical Upper limb Functional assessment

(1) Fugl-Meyer Assessment-Upper Extremity

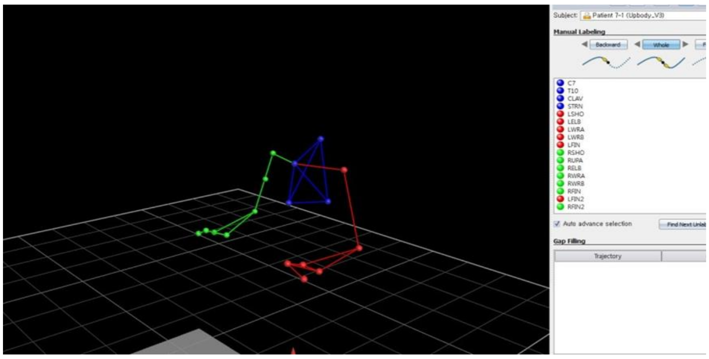
FMA-UE is a tool used to evaluate function of the shoulder, elbow, forearm, wrist, and hand in patients with stroke. It has high reliability, with test-retest reliability of 0.94 and inter-rater reliability of 0.99.[4-7] The FMA-UE score is based on multiple items measured on a three-point scale, with a maximum total score of 66 points. Overall upper limb functional impairment can be summarized in three categories: severe (0–28 points), moderate (29–58 points), and mild (59–66 points).[26-27]

(2) Box and Block Test

The BBT is an evaluation tool in which hand dexterity is measured based on the number of blocks a patient can move, one at a time, into a box in 1 min.[6] The test-retest reliability is 0.98, and the inter-rater reliability is 0.95, indicating high reliability. Task-oriented function of the upper limb can be summarized by the following equation: *Patient's Score – Mean Score / Standard Deviation (SD)*, where the Mean Score and SD refer to age- and gender-matched healthy individuals, and the result is classified as normal (0 to –2SD), mild (–2SD to –3SD), or severe (< –3SD).[28]

B. Three-dimensional upper limb motion analysis test

Upper limb motion analysis was performed by using a computerized 3D motion capture system (VICON MX-T10 Motion Analysis System, Oxford Metrics Inc., Oxford, UK) to record trajectories of reflective markers while patients performed the Reach & Grasp Cycle (Figure 2).[29-31] Patients performed the Reach & Grasp Cycle at a self-selected speed while sitting in front of a table in the motion analysis lab at Severance Rehabilitation Hospital. During the examination, 16 markers (C7, T10, clavicle, sternum, acromio-clavicle joint, lateral epicondyles, styloid processes of radius, heads of ulna, 2nd and 5th metacarpal joints of hands) were attached to both arms and the trunk according to the plug-in gait upper body model (Figure 3). From a starting position sitting at the table with the elbow and knee flexed at 90 degrees, the Reach & Grasp Cycle consists of four tasks: reaching for a cup on the table (T1), holding the cup and bringing it to the mouth (T2), putting the cup back in place (T3), and returning to the starting position (T4; Figure 4). The movements of each marker were recorded with six infrared cameras, and the coordinate data of each marker were obtained by post-processing on a computer with Nexus software version 1.8.5 connected to the motion analysis equipment. In addition, inverse kinematic analysis was performed to calculate the angles of each joint of the upper limb and determine any deficiency of upper limb movement.



Subject: Patient 7-1 (Upbody_V3)

☑ Auto advance selection

Figure 2. 3D upper limb motion capture

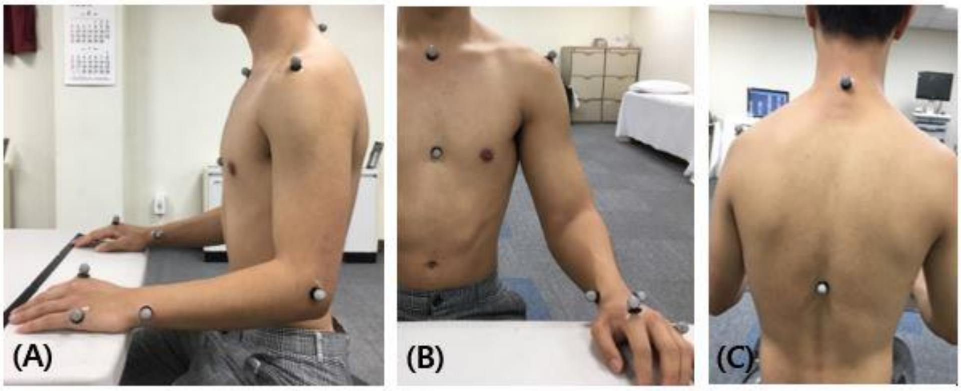


Figure 3. Marker set of 3D upper limb motion capture.

(A) Side view, (B) Front view, (c) Back view.

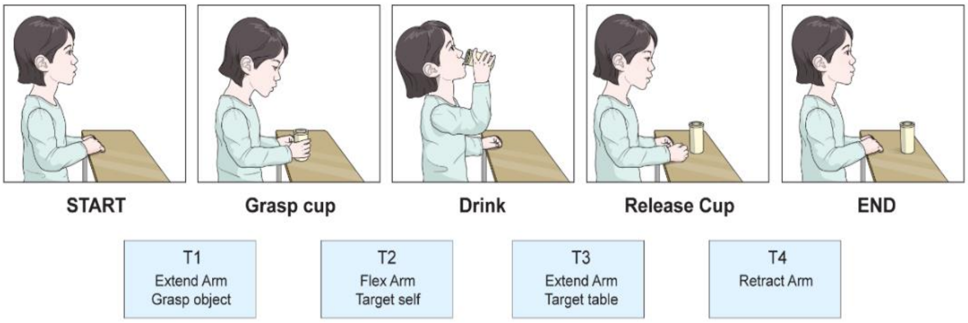


Figure 4. Reach & Grasp Cycle

C. Two-dimensional video

A 2D sagittal view of patients performing the Reach & Grasp Cycle during the 3D motion analysis test was recorded at 30 frames per second with a resolution of 1920 × 1080 pixels using a digital RGB camera positioned about 2–3 m from the patient’s seat (Figure 5).

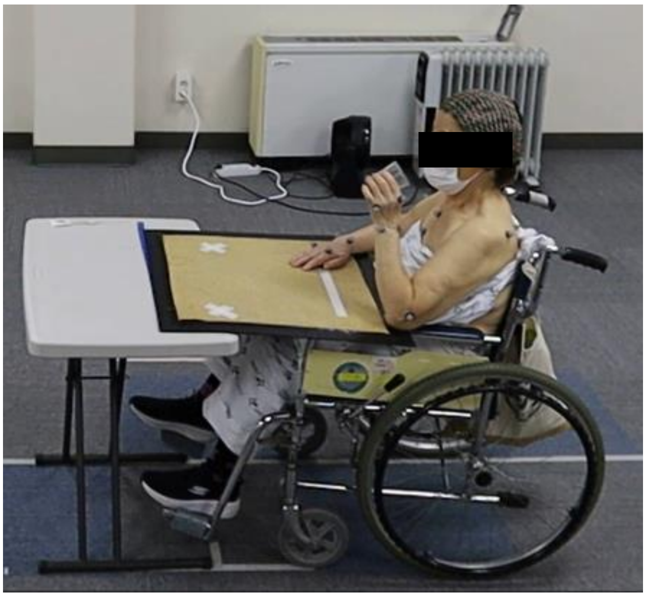

[redacted]

Figure 5. Sagittal view of 2D video

D. Metrics representing upper limb function

The variables representing upper limb function to be predicted using AI can be classified into two types as follows: 1) clinical metrics measured by clinician observation of patients and 2) parameters derived from the 3D motion analysis. The two clinical metrics used were the FMA-UE score and the BBT score, both of which can be categorized into three groups based on the severity of upper limb dysfunction. The parameters derived from the 3D motion analysis consisted of three spatiotemporal parameters (movement time, index of curvature, and number of movement units) and the Arm Profile Score (APS).[10-11] Movement time is the time required to complete each phase of the Reach & Grasp Cycle. Index of curvature, which represents the efficiency of upper limb movement, is calculated by dividing the length of the trajectory of the wrist marker during each phase of the Reach & Grasp Cycle by the linear distance between the initial and final marker positions (Figure 6). Number of movement units is a value representing the smoothness of upper limb movement and is calculated by calculating the number of acceleration–deceleration inflection points in the velocity profile of the wrist marker during the Reach & Grasp Cycle (Figure 7). The APS is a kinematic parameter calculated from 3D motion capture data by determining the Root Mean Square Error (RMSE) value between the kinematic data of individuals with upper limb dysfunction and the average kinematic data of individuals without upper limb pathology (Figure 8).[11] Specifically, the APS is an average of 10 Arm Variable Scores: Trunk Tilt, Trunk Obliquity, Trunk Rotation, Shoulder Flexion/Extension, Shoulder Abduction/Adduction, Shoulder Rotation, Elbow Flexion/Extension, Wrist Flexion/Extension, Wrist Deviation, and Wrist Rotation. The higher the APS, the higher the severity of upper limb movement impairment.

We classified clinical function evaluation FMA-UE and BBT into 3 groups, and estimated the parameter derived from 3D motion capture data through regression with the coordinate data obtained from 3D motion capture using the deep learning model (Table 1).

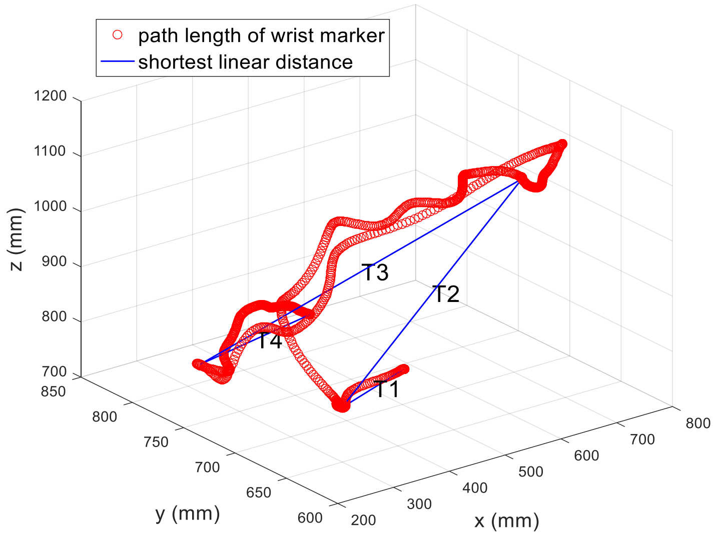


Figure 6. Index of curvature. The path length of wrist marker divided by shortest linear distance during each phase of the Reach & Grasp Cycle

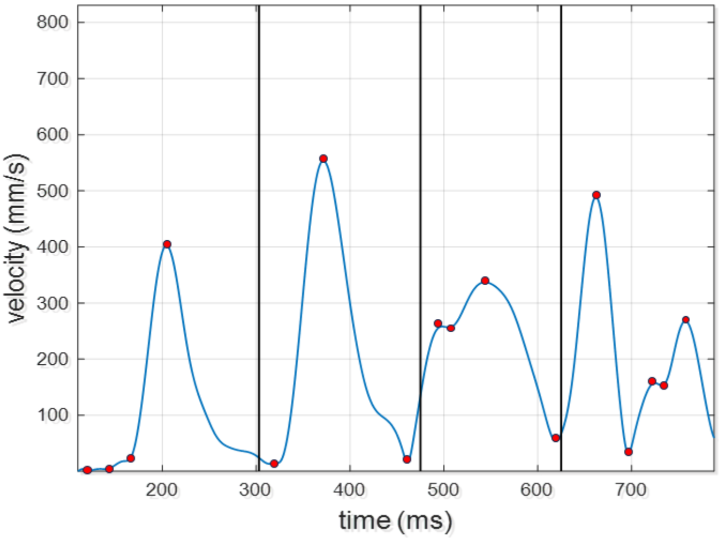


Figure 7. Number of movement units. The number of red circles: acceleration–deceleration inflection points in the velocity profile of the wrist marker during the Reach & Grasp Cycle.

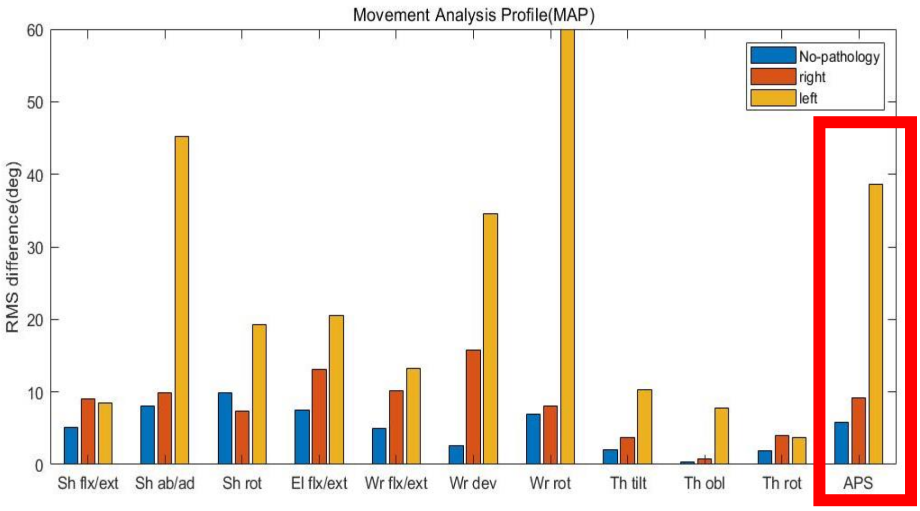


Figure 8. Arm Profile Score. The average of root mean square error values of 10 upper limb movements.

Table 1. Parameters predicted by deep learning

	Parameters		Prediction method
Clinical metrics	1. Fugl-Meyer Assessment - Upper Extremity		Classification of 3 groups (severe, moderate, mild)
	2. Box and Block Test		Classification of 3 groups (severe, mild, normal)
Parameters derived from the 3D motion analysis	Temporospatial parameters	3. Movement times	Regression of a continuous value
		4. Index of curvature	
		5. Number of movement units	
	Kinematic parameter	6. Arm Profile Score	Regression of a continuous value

E. Input data

(1) Three-dimensional motion capture dataset

The input data for AI to predict upper limb function metrics based on 3D motion capture data consisted of coordinate values of 10 markers: Trunk (4), Shoulder, Elbow, Wrist (2), and Finger (2). To solve the problem of global translation, the coordinate value of the sternum marker was set to 0 for each patient. In addition, because each patient performed the Reach & Grasp Cycle at a self-selected speed, time normalization was performed using TimeSeriesResampler to set all data frames to 2000. The format of the 3D coordinate input data (relative X, Y and Z coordinates of the 10 markers) represented a 2D matrix with a feature dimension (30) as a vertical axis and a time dimension (2000) as a horizontal axis (Figure 9). The 3D motion capture data included 624 datasets of 265 patients. To train the deep learning models, these datasets were divided into a training set (60%), a validation set (20%), and a test set (20%; Table 2).

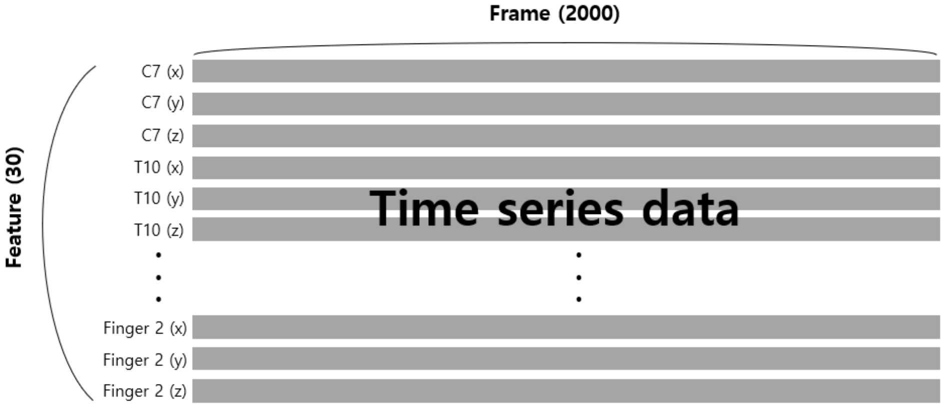


Figure 9. Input dataset format 1. Time series data of 3D coordinates measured by 3D motion capture

Table 2. Number of 3D coordinate dataset

	Train	Validation	Test	Total
Fugl-Meyer Assessment -Upper Extremity	373	127	124	624
Box and Block Test	370	127	124	621
Arm Profile Score	347	114	114	575
Temporospatial parameters	347	114	114	575

(2) Two-dimensional video dataset

Two-dimensional keypoints recorded in 2D video while patients performed the Reach & Grasp Cycle were extracted using the RTMPose algorithm applied through the MMPose tool developed by Open-mmlab (Figure 10). RTMPose follows a top-down paradigm by first finding the object bounding box using CSPNeXt as a backbone model, which has excellent speed and accuracy, and then estimating each pose individually using a SimCC-based algorithm, which has competitive accuracy with relatively few calculations.[33-35] Before it was applied to the video of patients performing the Reach & Grasp Cycle, RTMPose was trained with the COCO-WholeBody dataset, which contains annotated whole-body keypoints from 200,000 images.[32] The COCO-WholeBody dataset is an extension of the COCO dataset and includes a total of 133 keypoints, with 68 detailed keypoints on the face, 42 on the hand, and 6 on the foot added to 17 existing keypoints for the body (Figure 11). The COCO-WholeBody dataset, rather than the more commonly used COCO dataset, was used because it contains detailed hand keypoints important in upper limb movements. The 2D keypoints in the COCO-WholeBody dataset used to train RTMPose were similar to the markers used in the 3D motion capture dataset and included the shoulder (6, 7), elbow (8, 9), wrist (10, 11), carpometacarpal joint of the thumb (93, 113), and metacarpal joints of the 2nd and 5th fingers (97, 109, 118, 130). The median value of the right shoulder and left shoulder was used for the trunk.

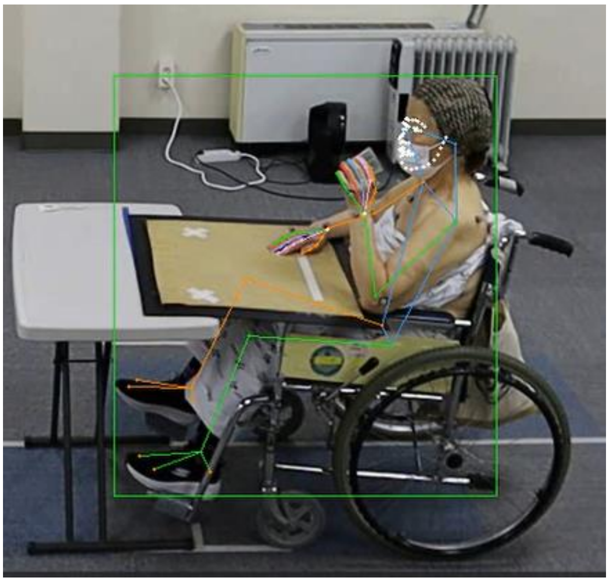


Figure 10. Pose estimation by RTMPose. A green box: the box that detects the subject, Small circles and the lines: keypoints of the subject and the lines connecting keypoints.

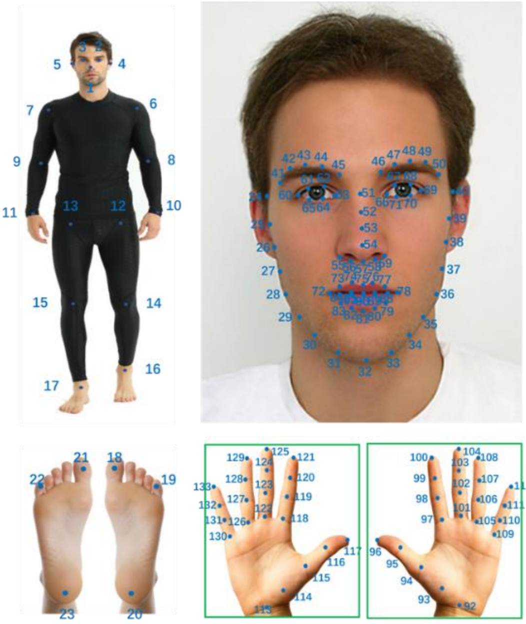


Figure 11. Keypoint annotations of COCO-wholebody dataset

The coordinate values of seven keypoints in the 2D video dataset [Trunk, Shoulder, Elbow, Wrist (2), and Finger (2)] were used as input data for predicting upper limb function metrics with AI. To solve the problem of global translation, normalization was performed to the maximum size of the bounding box that recognizes individuals in the 2D video. The X and Y coordinates of the exact center of the bounding box were each set to 0, and the maximum and minimum values of the coordinates of each keypoint were set to 1 and –1, respectively. In the 2D video dataset, the horizontal position was indicated on the X axis, and the vertical position was indicated on the Y axis. In addition, because each patient performed the Reach & Grasp Cycle at a self-selected speed, time normalization was performed using TimeSeriesResampler to set all data frames to 600. The format of the 2D video input data (relative X and Y coordinates of seven keypoints) represented a 2D matrix with a feature dimension (14) as a vertical axis and a time dimension (600) as a horizontal axis (Figure 12). The 2D video data included 330 datasets of 103 patients. To train the deep learning model, the input data were divided into a training set (60%), a validation set (20%), and a test set (20%; Table 3).

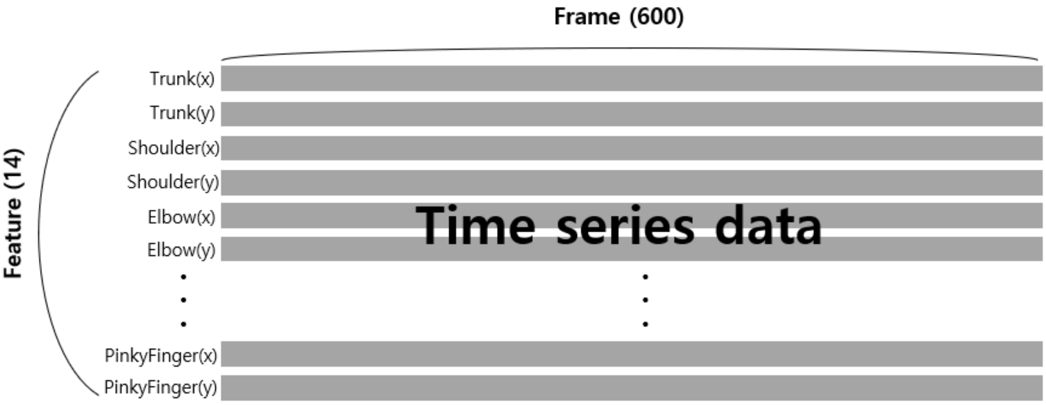


Figure 12. Input dataset format 2. Time series data of 2D keypoints estimated in 2D videos.

Table 3. Number of 2D keypoint dataset

	Train	Validation	Test	Total
Fugl-Meyer Assessment -Upper Extremity	199	66	65	330
Box and Block Test	199	66	65	330
Arm Profile Score	117	39	40	196
Temporospatial parameters	117	39	40	196

4. Deep learning model

A. Convolutional Neural Network

A Convolutional Neural Network (CNN) is a deep learning algorithm that excels in image and time series data recognition and processing. The core of the CNN approach is to extract and understand the data features through convolution and pooling. Convolution refers to the process of extracting information by moving a small filter over the data. During this process, the filter is optimized to recognize a specific pattern, giving the CNN the ability to extract high-level information by recognizing various features. Pooling is the process of simplifying the information extracted from convolutions. By reducing the size of the data or emphasizing certain information, pooling enables efficient processing by leaving only notable features. A one-dimensional (1D) CNN model moves in only one direction in sequence data, making it very suitable for time series analysis. Therefore, we used a 1D CNN model to detect and train patterns according to spatial dimensions.

(1) Three-dimensional motion capture dataset

We classified 3D datasets according to predicted FMA-UE and BBT scores by training a CNN model with time series coordinate data from 3D motion capture of the Reach & Grasp Cycle. The CNN architecture for classification is shown in Figure 13. Convolution was performed with a filter size of 3 and a stride of 1. To overcome the vanishing gradient problem, we used leaky rectified linear unit (leakyRelu) as the activation function, which allows models to learn faster and perform more efficiently. After convolution, max pooling was performed to simplify the time axis data. The CNN was trained with a total of six convolution layers + pooling layers. The convolution block was composed of 1D convolutional layers, which are widely used in time series data processing because they calculate the output while moving only horizontally. Adam was used as the optimizer, and the loss function minimized by the model was cross-entropy. The batch size was 64 and the initial learning rate was 0.001. Drop-out was set to 0.4 to randomly remove some neurons to prevent the model from becoming overly dependent on specific neurons. Iteratively convolutioned and pooled data were flattened to create a fully connected neural network, and the softmax function was used to find the probability of belonging to each of the three categories of FMA-UE and BBT.

We also used a CNN model to estimate continuous values of spatiotemporal parameters and APS in 3D datasets by regression. The CNN architecture for regression is shown in Figure 14. The CNN regression model was trained with the same 1D CNN architecture as the classification model with three stacked convolution layers. The loss function was the mean square error (MSE), and the optimizer was Adam. The batch size was 64 and the initial learning rate was 0.001. After flattening, the output was changed to 1 so that a continuous value came out.

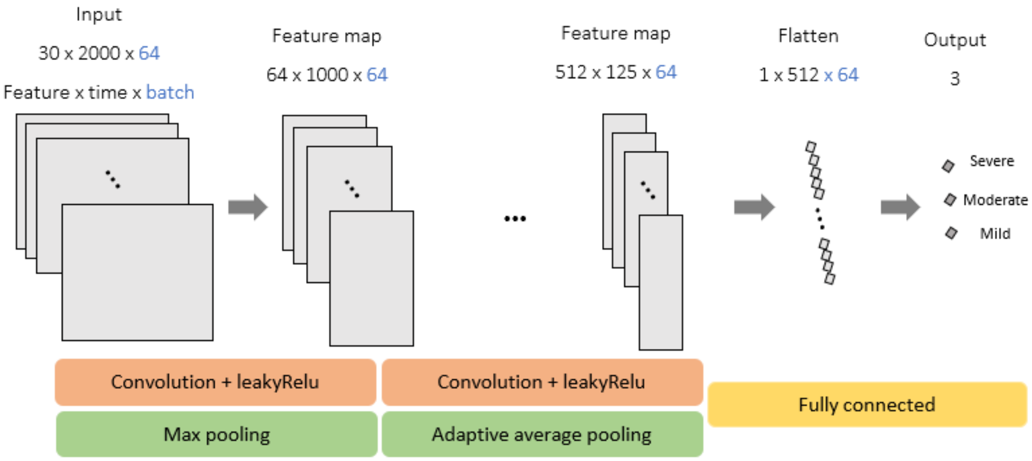


Figure 13. The architecture of the CNN classification using 3D motion capture dataset

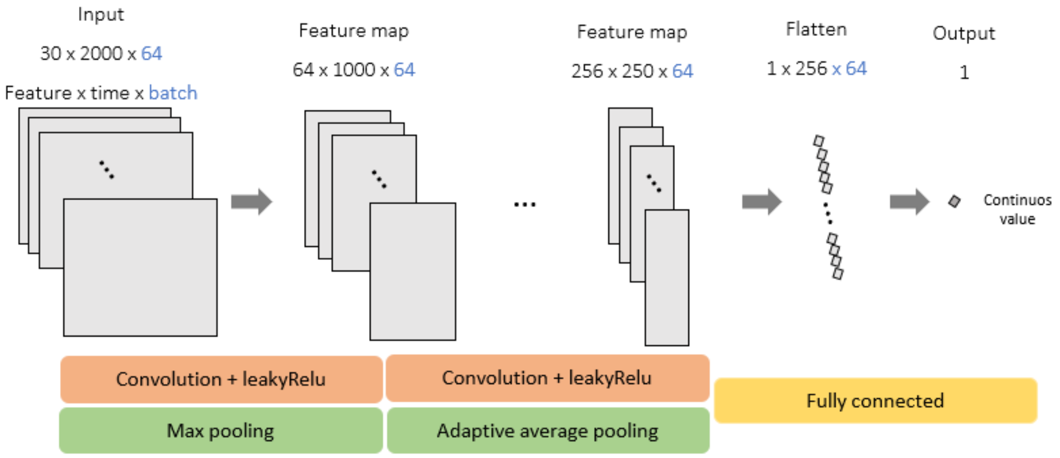


Figure 14. The architecture of the CNN regression using 3D motion capture dataset

(2) Two-dimensional video dataset

We classified 2D datasets according to predicted FMA-UE and BBT scores by training a CNN model with time series keypoint coordinate data from 2D video of the Reach & Grab Cycle. The CNN architecture for classification is shown in Figure 15. Convolution was performed with a filter size of 7 and a stride of 1, with leakyRelu as the activation function. After convolution, max pooling was performed to simplify the time axis data. The model was trained with a total of six convolution layers + pooling layers. The convolution block was composed of 1D convolutional layers. Adam was used as the optimizer, and the loss function was cross-entropy. The batch size was 32, and the initial learning rate was 0.0001. Drop-out was set to 0.3.

We also used a CNN model to estimate continuous values of APS and temporospatial parameters in 2D datasets by regression. The CNN architecture for regression is shown in Figure 16. The CNN regression model was trained with a 1D CNN architecture in the same way as the classification model with seven stacked convolution layers. The loss function was MSE, and the optimizer was Adam. The batch size was 64, and the initial learning rate was 0.001. After flattening, the output was changed to 1 so that continuous values came out.

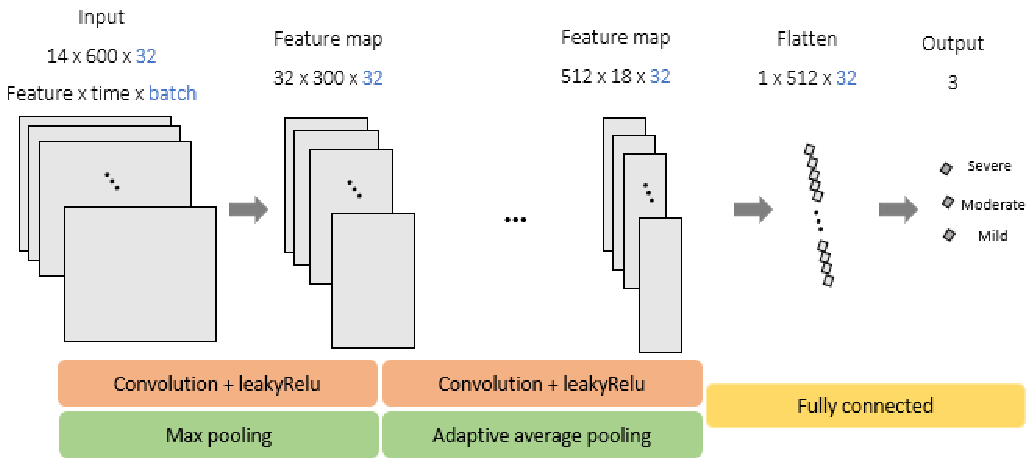


Figure 15. The architecture of the CNN classification using 2D video dataset

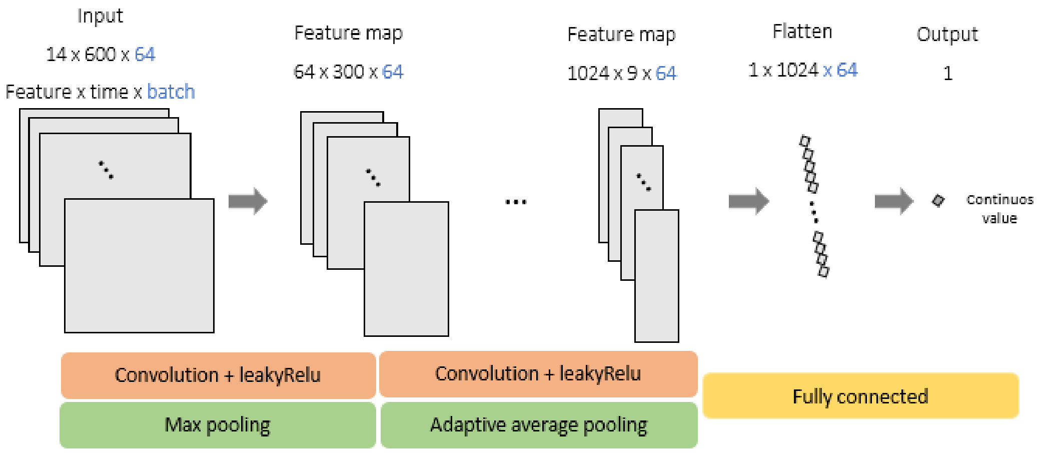


Figure 16. The architecture of the CNN regression using 2D video dataset

B. Temporal Convolutional Network

A Temporal Convolutional Network (TCN) is a deep learning model that shows high performance with time series datasets. Because of the characteristics of time series data, a TCN must satisfy causality, which depends on only the present and the past and not the future. A TCN uses causal convolution to prevent information from the future from flowing into the output at the current time. Additionally, dilated convolution is used to efficiently identify information from distant time steps by 1D convolution. By using dilated convolution to add zero padding inside the filter, the receptive field is increased and calculations are reduced. By using causal dilated convolutional layers, a TCN reduces spatial dimension loss and improves computational efficiency by increasing the receptive field without using pooling in the sequence (Figure 17). Figure 18 shows the overall architecture of the TCN used in this study.

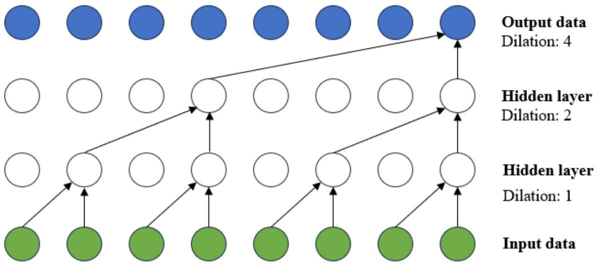


Figure 17. Causal dilated convolutional layers of TCN

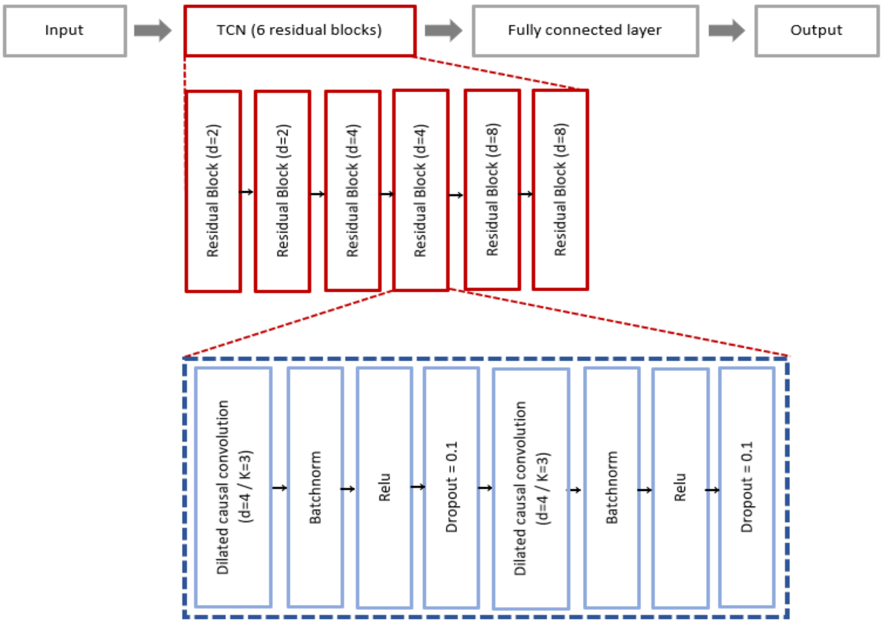


Figure 18. The architecture of TCN

(1) Three-dimensional motion capture dataset

We classified 3D datasets according to predicted FMA-UE and BBT scores by training the TCN model with the same 3D time series coordinate data used to train the CNN model. The residual block was composed of the dilated causal convolutional layer, the batchnorm layer, the activation layer, and the drop-out layer. For dilated causal convolution, the kernel size was 3, and 2, 4, and 8 were each used twice as dilated factors. Relu was used as the activation function, and Adam was used as the optimizer. The loss function was cross-entropy. The batch size was 32, and the initial learning rate was 0.0001. Iteratively dilated convolutional data were flattened to create a fully connected neural network. The softmax function was used to find the probability of datasets belonging to each of the three categories of FMA-UE and BBT scores.

We also estimated continuous values of APS and temporospatial parameters in 3D datasets by regression using a TCN model. The TCN regression model was trained with the same architecture as the TCN classification model. The loss function was MSE, and the optimizer was Adam. After flattening, the output was changed to 1 so that continuous values came out.

(2) Two-dimensional video dataset

We classified datasets according to predicted FMA-UE and BBT scores by training the TCN model with time series keypoint coordinate data from 2D video of the Reach & Grab Cycle. The model was composed of the same TCN architecture used for the 3D motion capture data. The initial learning rate was set to 0.1 to address overfitting problems related to the small dataset. We also estimated continuous values of APS and temporospatial parameters in 2D datasets by regression using a TCN model. The model used had the same architecture as the TCN model used for the 3D motion capture data.

5. Evaluation of performance

We calculated the accuracy, precision, recall, and F1-scores of the trained AI models to evaluate the models' performance. The accuracy is the total number of correct predictions divided by the total number of matched ground-truth values. The precision was calculated as the ratio of ground-truth positives to predicted positives. The recall was calculated as the ratio of predicted positives to ground-truth positives. The F1-score is an index defined as the harmonic average of the precision and the recall. To evaluate regression performance, we calculated the correlation coefficient between predicted values and ground-truth values.

III. RESULTS

1. Three-dimensional motion capture dataset

A. FMA-UE classification

We used the CNN and TCN models with data to classify 3D motion capture datasets into three categories based on predicted FMA-UE scores. Figure 19 shows the area under the receiver operating characteristic curve (ROC-AUC) for each category, the values of which were 0.98 (severe), 0.96 (moderate), and 0.97 (mild) for the CNN model and 0.95, 0.57, and 0.92 for the TCN model, respectively. The results of FMA-UE classification were also shown in the confusion matrix (Figure 20). Table 4 summarizes the performance of the deep learning models for FMA-UE classification using the 3D motion capture data. The accuracy, precision, recall, and F1-score of FMA-UE classification were 91.13%, 90.27%, 90.35%, and 90.31% for the CNN model and 79.03%, 72.54%, 68.17%, and 73.19% for the TCN model, respectively.

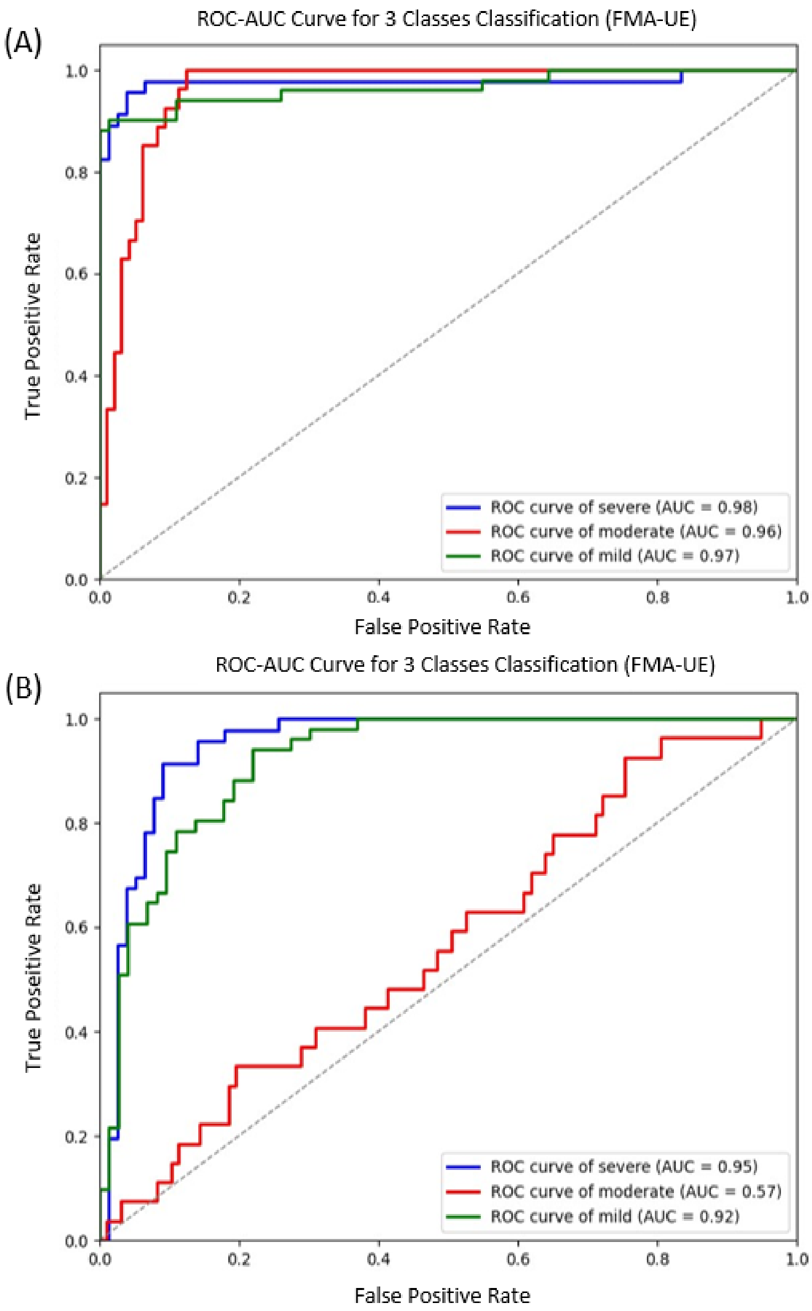


Figure 19. The ROC-AUC of the FMA-UE classification using 3D motion capture dataset. (A) CNN, (B) TCN.

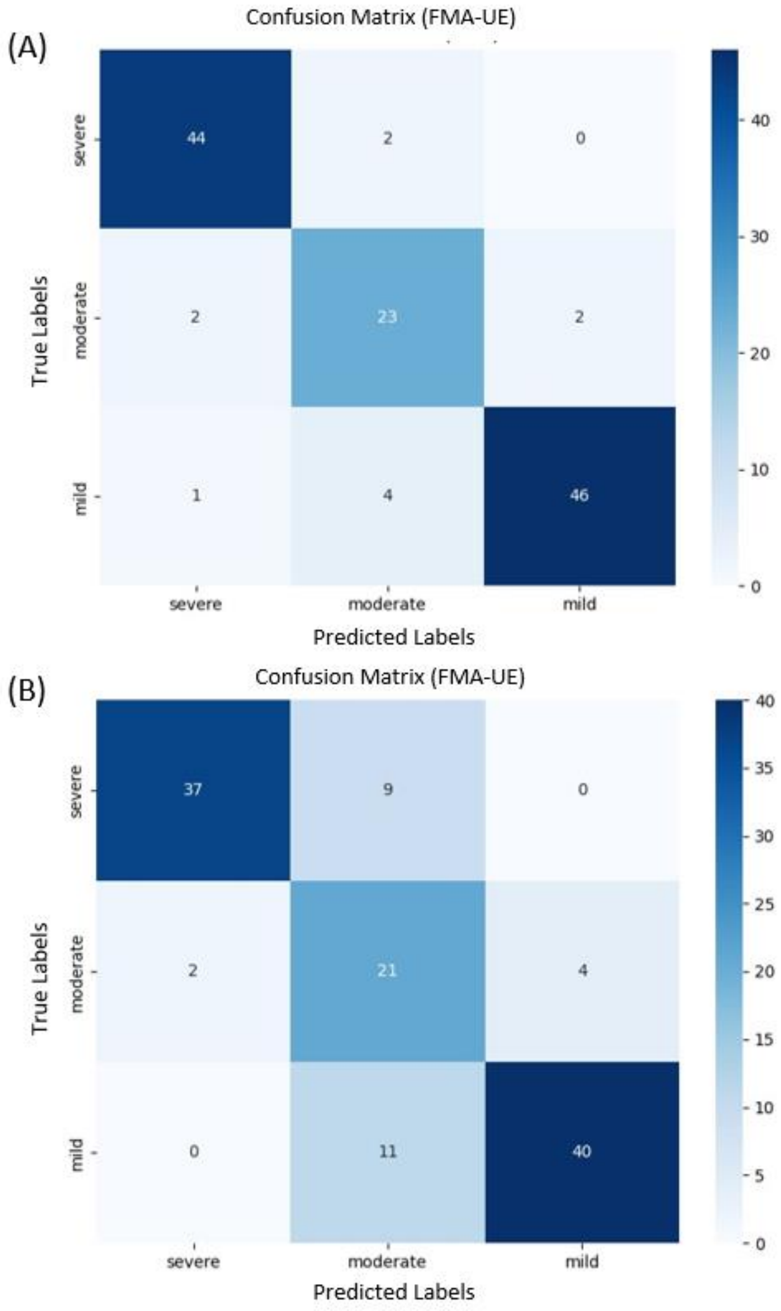


Figure 20. The confusion matrices of the FMA classification using 3D motion capture dataset. (A) CNN, (B) TCN.

B. BBT classification

We used the CNN and TCN models to classify 3D motion capture datasets into three categories based on predicted BBT scores. Figure 21 shows the ROC-AUC for each category, the values of which were 0.94 (severe), 0.82 (mild), and 0.86 (normal) for the CNN model and 0.92 (severe), 0.62 (mild), and 0.85 (normal) for the TCN model. The results of BTT classification are also shown in the confusion matrix (Figure 22). The performance of the CNN and TCN models for BBT classification is summarized in Table 4. The accuracy, precision, recall, and F1-score of BBT classification were 79.03%, 72.54%, 73.96%, and 73.24% for the CNN model and 75.81%, 66.87%, 65.06%, and 65.95% for the TCN model, respectively.

Table 4. Performance of classification using 3D motion capture dataset

	Model	Group	Accuracy	Precision	Recall	F1-score
Fugl-Meyer Assessment -Upper Extremity	CNN	Severe	-	95.65	95.65	95.64
		Moderate	-	79.31	85.19	82.14
		Mild	-	95.83	90.20	92.93
		Total	91.13	90.27	90.35	90.31
	TCN	Severe	-	94.87	80.43	87.06
		Moderate	-	51.22	45.65	48.27
		Mild	-	90.91	78.43	84.17
		Total	79.03	79.00	68.17	73.19
Box and Block Test	CNN	Severe	-	95.31	88.41	91.73
		Mild	-	53.57	62.50	57.69
		Normal	-	68.75	70.97	70.40
		Total	79.03	72.54	73.96	73.24
	TCN	Severe	-	83.95	98.55	90.67
		Mild	-	50.00	33.33	40.00
		Normal	-	66.67	63.31	62.07
		Total	75.81	66.87	65.06	65.95

[1]CNN, Convolutional Neural Network; TCN, Temporal Convolutional Network

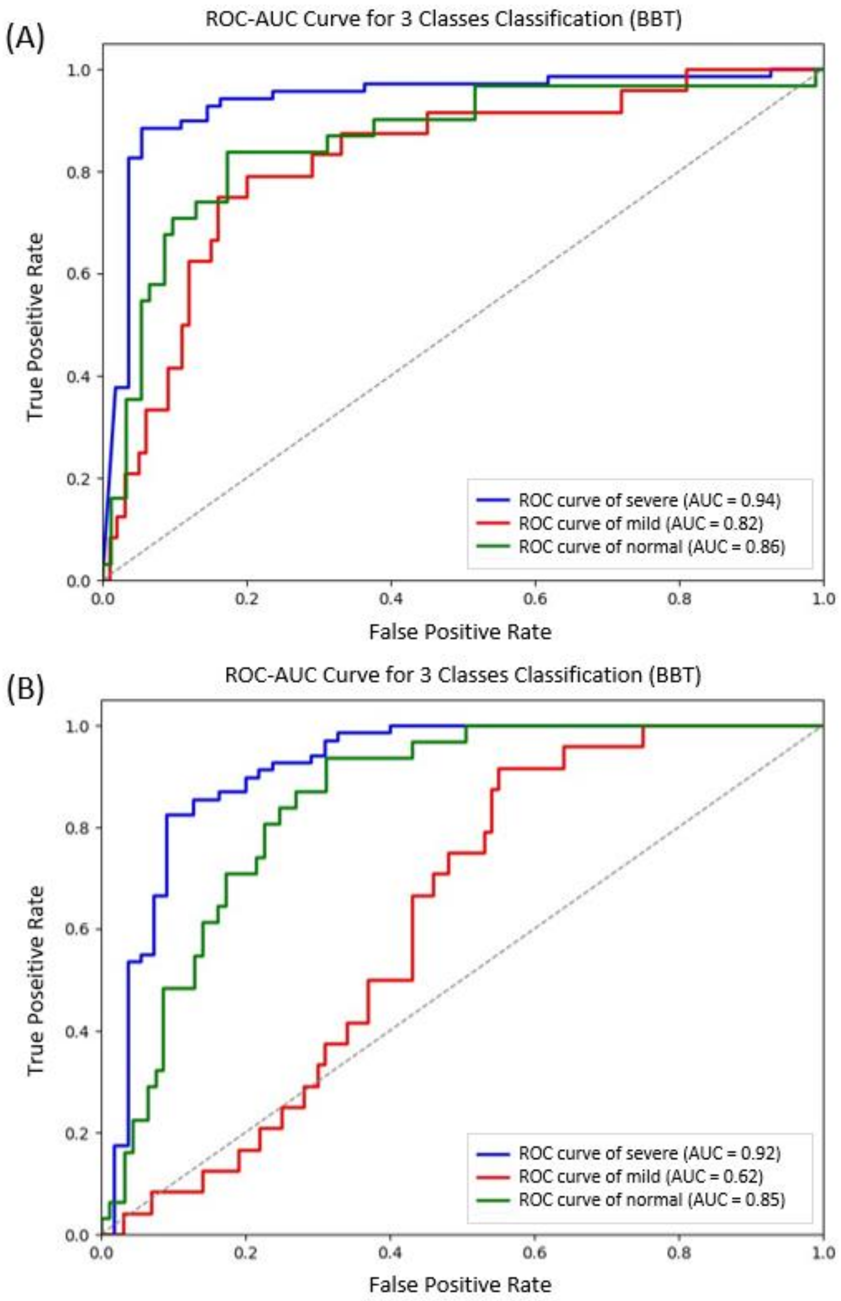


Figure 21. The ROC-AUC of the BBT classification using 3D motion capture dataset. (A) CNN, (B) TCN.

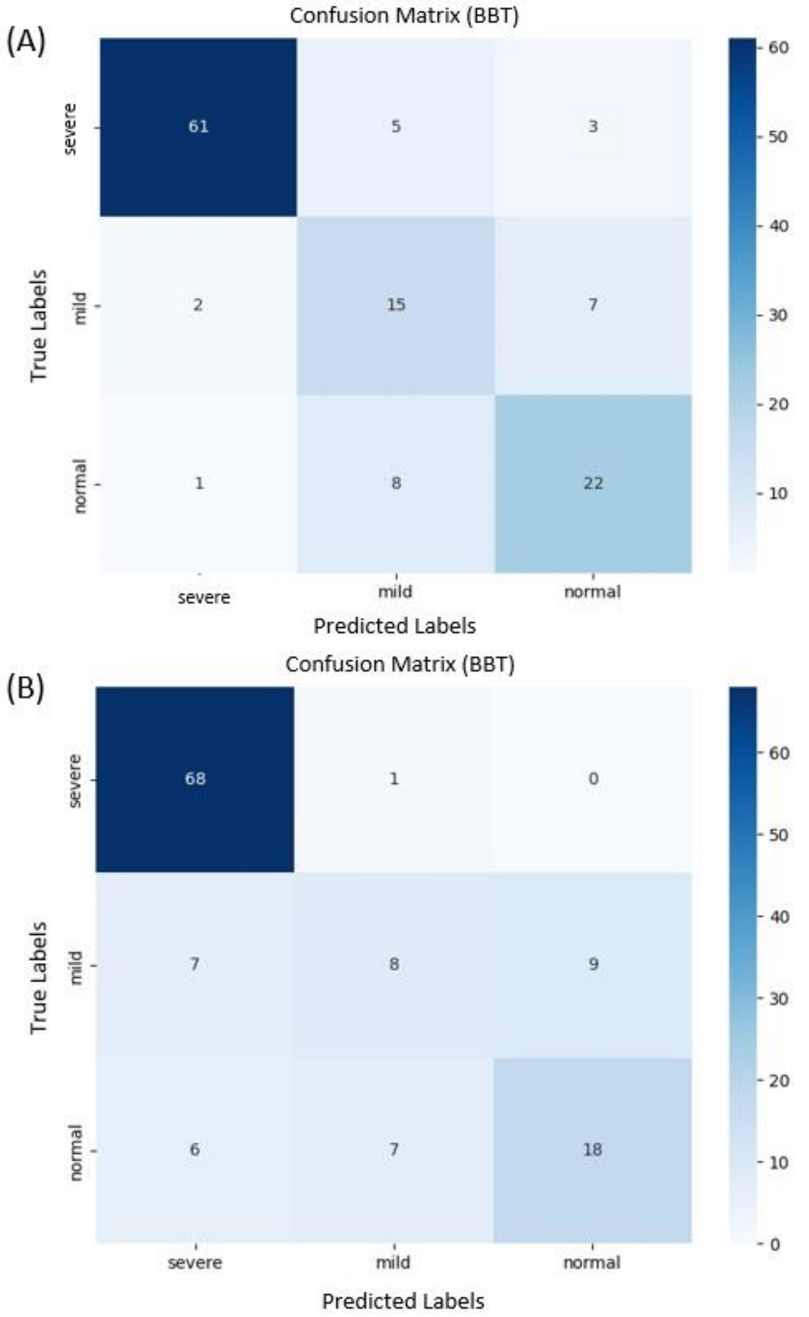


Figure 22. The confusion matrices of the BBT classification using 3D motion capture dataset. (A) CNN, (B) TCN.

C. APS regression

We used the CNN and TCN models to predict APS using the 3D coordinate data obtained from 3D motion capture. Table 5 shows the performance of APS regression using the 3D motion capture data. The correlation coefficient between the predicted APS and the ground-truth data obtained by 3D motion capture was 0.783 for the CNN model and 0.668 for the TCN model (Figure 23). The mean error value between the predicted and ground-truth values was 5.46 for the CNN model and –0.93 for the TCN model (Figure 24).

D. Temporospatial parameters regression

We used the CNN and TCN models to predict movement time, index of curvature, and number of movement units in the 3D coordinate datasets. Table 5 shows the performance of temporospatial parameter regression using the 3D motion capture data. The correlation coefficients between the predicted and ground-truth values of movement time, index of curvature, and number of movement units were 0.544, 0.755, and 0.601 for the CNN model and 0.399, 0.531, and 0.408 for the TCN model (Figure 23). The mean errors between the predicted and ground-truth values of movement time, index of curvature, and number of movement units were 2.84, 1.68, and 5.23 for the CNN model and –0.29, –0.23, and –0.85 for the TCN model, respectively (Figure 24).

Table 5. Performance of regression using 3D motion capture dataset

	Model	Mean error [95% confidence interval]	Correlation coefficient
Arm Profile Score	CNN	5.46 [-2.36, 13.28]	0.783*
	TCN	-0.93 [-8.61, 6.75]	0.668*
Movement times	CNN	2.84 [-14.55, 13.14]	0.544*
	TCN	-0.29 [-7.56, 6.98]	0.399*
Index of curvature	CNN	1.68 [0.91, 2.87]	0.755*
	TCN	-0.23 [-1.81, 1.35]	0.531*
Number of movement units	CNN	5.23 [-10.29, 34.29]	0.601*
	TCN	-0.85 [-27.77, 22.04]	0.408*

[1]CNN, Convolutional Neural Network; TCN, Temporal Convolutional Network
* p < 0.01 by Pearson correlation test

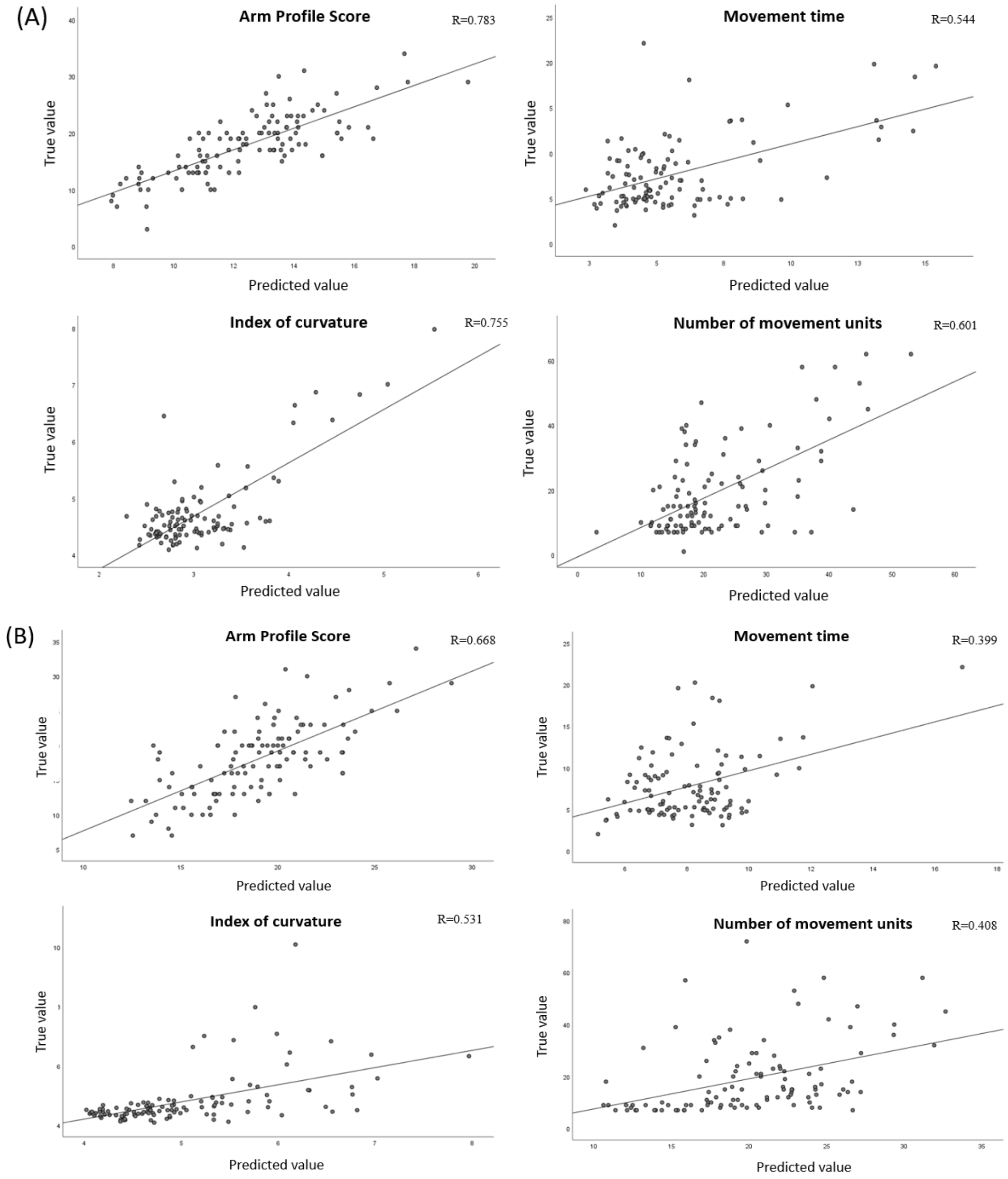


Figure 23. Scatter plots between true values and predicted values by regression using 3D motion capture dataset. (A) CNN, (B) TCN

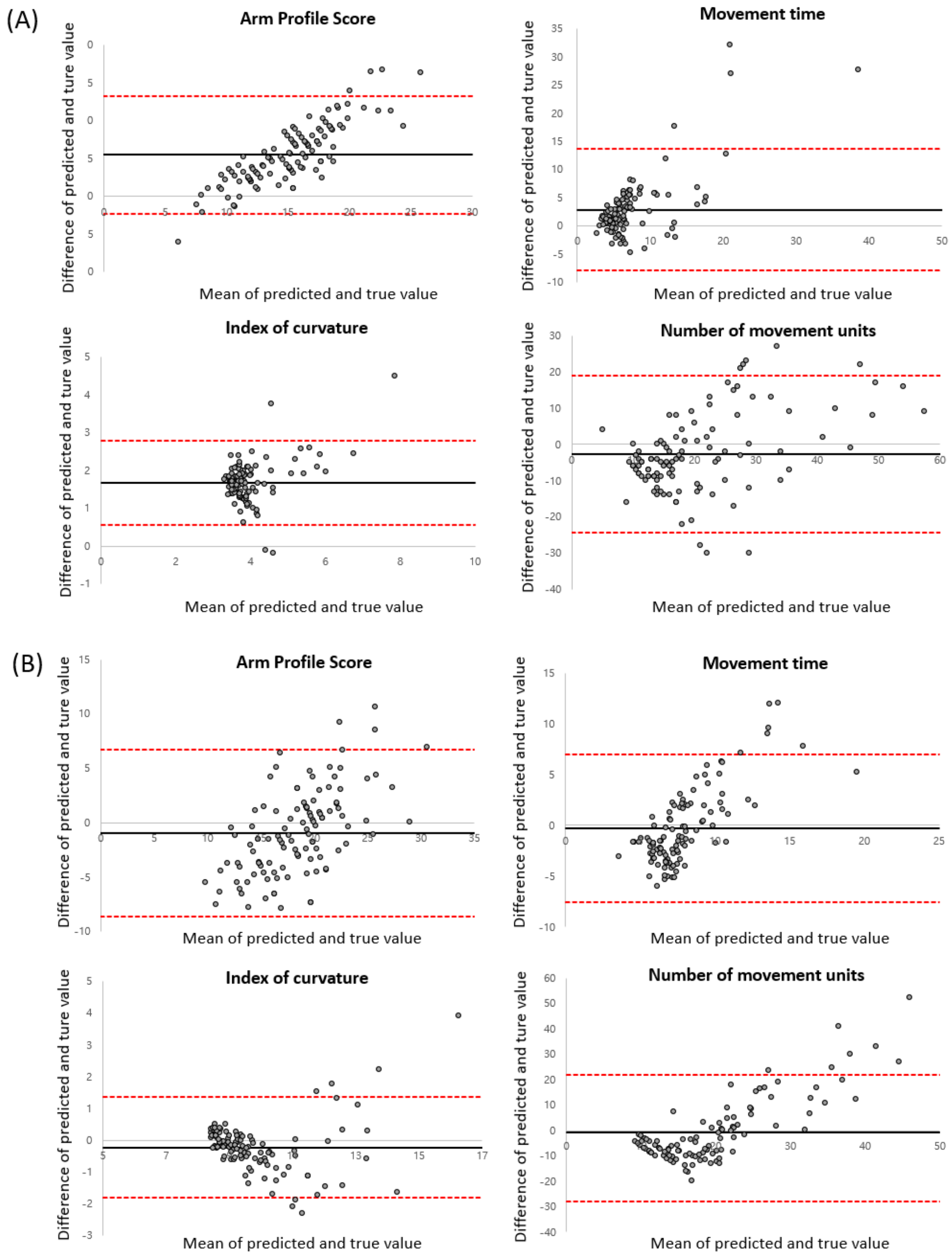


Figure 24. Bland-Altman plots between true values and predicted values by regression using 3D motion capture dataset. (A) CNN, (B) TCN. Black solid line: mean error, Red dotted line: 95% confidence interval.

2. Two-dimensional video dataset

A. FMA-UE classification

We used CNN and TCN models to classify 2D keypoint coordinate datasets into three categories according to predicted FMA-UE scores. Figure 25 shows the ROC-AUC for each category, the values of which were 0.96 (severe), 0.88 (moderate), and 0.96 (mild) for the CNN model and 0.93 (severe), 0.57 (moderate), and 0.81 (mild) for the TCN model. The results of FMA-UE classification using 2D data are also shown by the confusion matrix (Figure 26). Table 6 summarizes the performance of FMA-UE classification by the deep learning models using the 2D video data. The accuracy, precision, recall, and F1-score of FMA-UE classification were 89.23%, 88.39%, 85.97%, and 87.16% for the CNN model and 78.46%, 69.10%, 76.90%, and 74.73% for the TCN model, respectively.

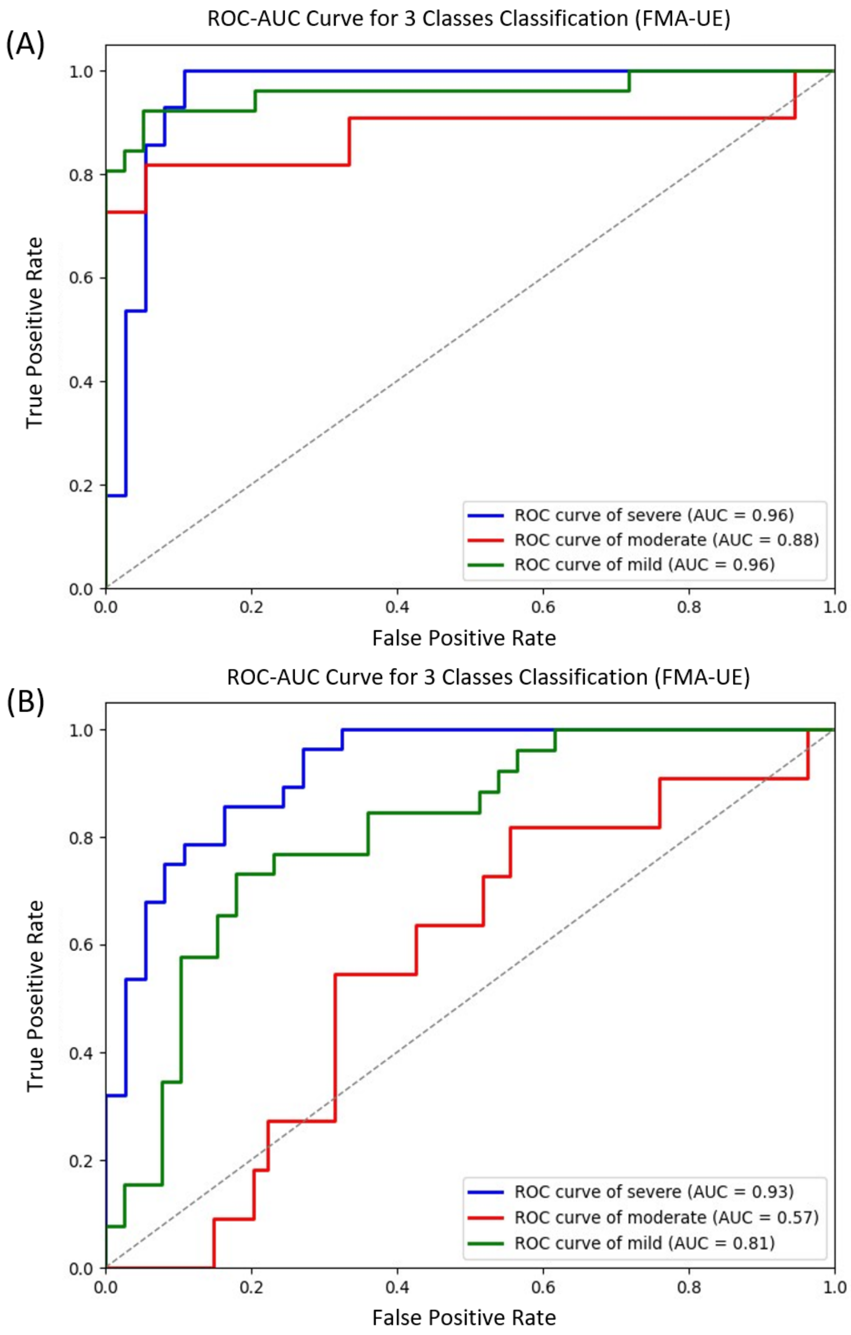


Figure 25. The ROC-AUC of the FMA classification using 2D video dataset. (A) CNN, (B) TCN.

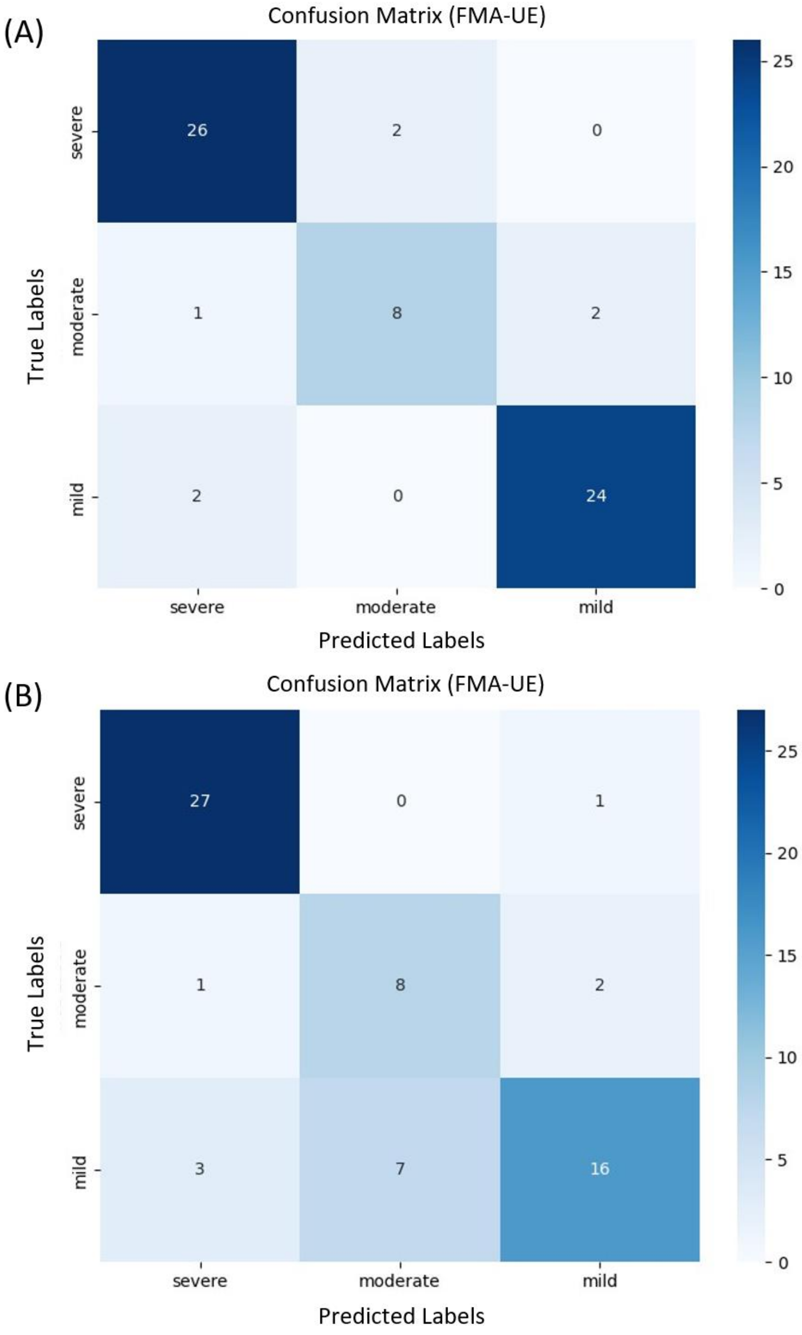


Figure 26. The confusion matrices of the FMA-UE classification using 2D video dataset. (A) CNN, (B) TCN.

B. BBT classification

We used the CNN and TCN models to classify the 2D keypoint coordinate datasets into three categories according to predicted BBT scores. Figure 27 shows the ROC-AUC for each category, the values of which were 0.90 (severe), 0.85 (mild), and 0.95 (normal) for the CNN model and 0.81 (severe), 0.70 (mild), and 0.73 (normal) for the TCN model. The results of BBT classification are also shown by the confusion matrix (Figure 28). The performance of the models for BBT classification with 2D video data is summarized in Table 6. The accuracy, precision, recall, and F1-score of BBT classification were 76.92%, 73.79%, 75.51%, and 74.64% for the CNN model and 70.77%, 67.81%, 67.40%, and 67.60% for the TCN model, respectively.

Table 6. Performance of classification using 2D video dataset

	Model	Group	Accuracy	Precision	Recall	F1-score
Fugl-Meyer Assessment -Upper Extremity	CNN	Severe	-	92.86	92.86	92.86
		Moderate	-	80.00	72.73	76.19
		Mild	-	92.31	92.31	92.31
		Total	89.23	88.39	85.97	87.16
	TCN	Severe	-	87.10	96.43	91.53
		Moderate	-	53.33	72.73	61.54
		Mild	-	84.21	61.54	71.11
		Total	78.46	69.10	76.90	74.73
Box and Block Test	CNN	Severe	-	90.00	81.82	85.72
		Mild	-	64.71	64.71	64.71
		Normal	-	66.67	80.00	72.73
		Total	76.92	73.79	75.51	74.64
	TCN	Severe	-	84.38	81.82	83.08
		Mild	-	66.67	47.06	55.17
		Normal	-	52.38	73.33	61.11
		Total	70.77	67.81	67.40	67.60

[1] CNN, Convolutional Neural Network; TCN, Temporal Convolutional Network

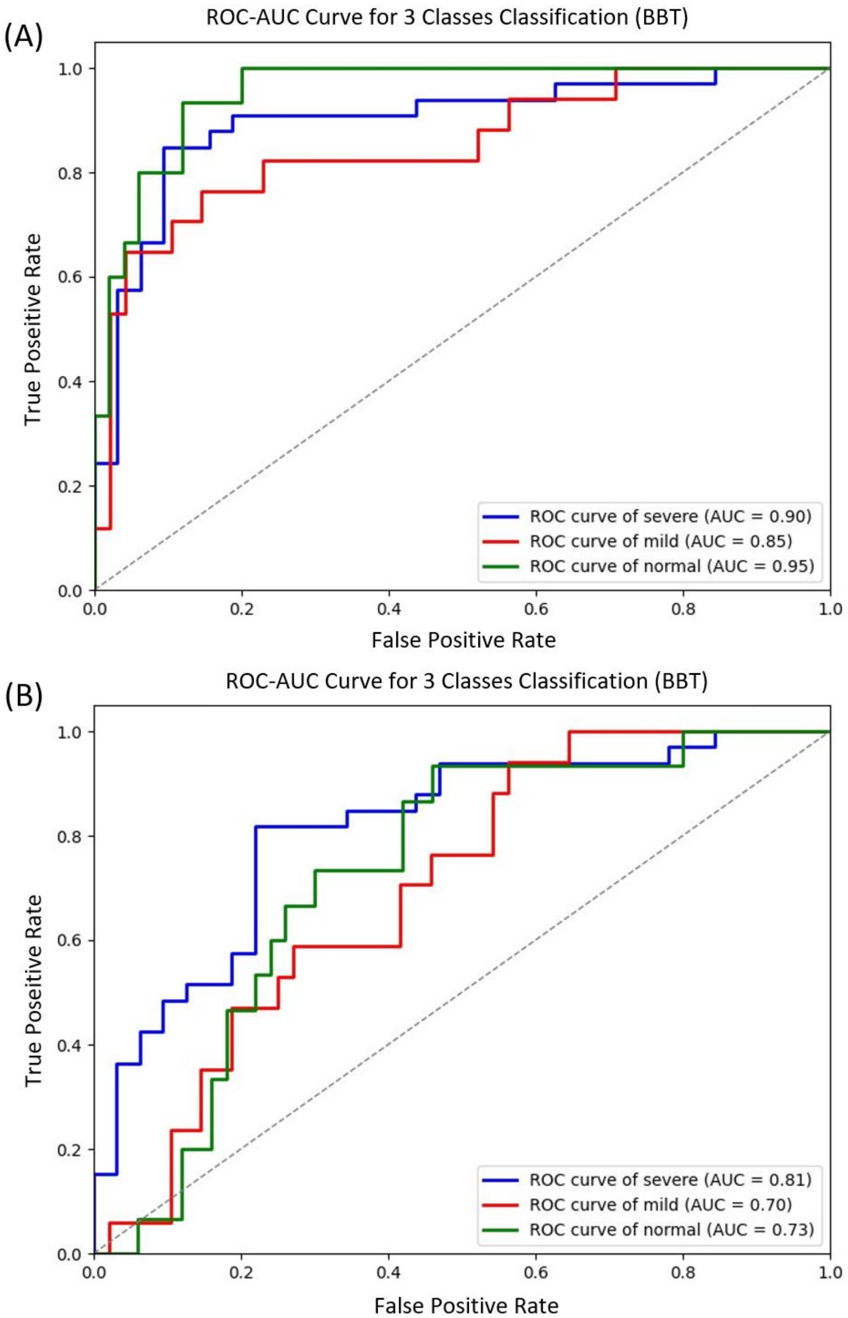


Figure 27. The ROC-AUC of the FMA classification using 2D video dataset. (A) CNN, (B) TCN.

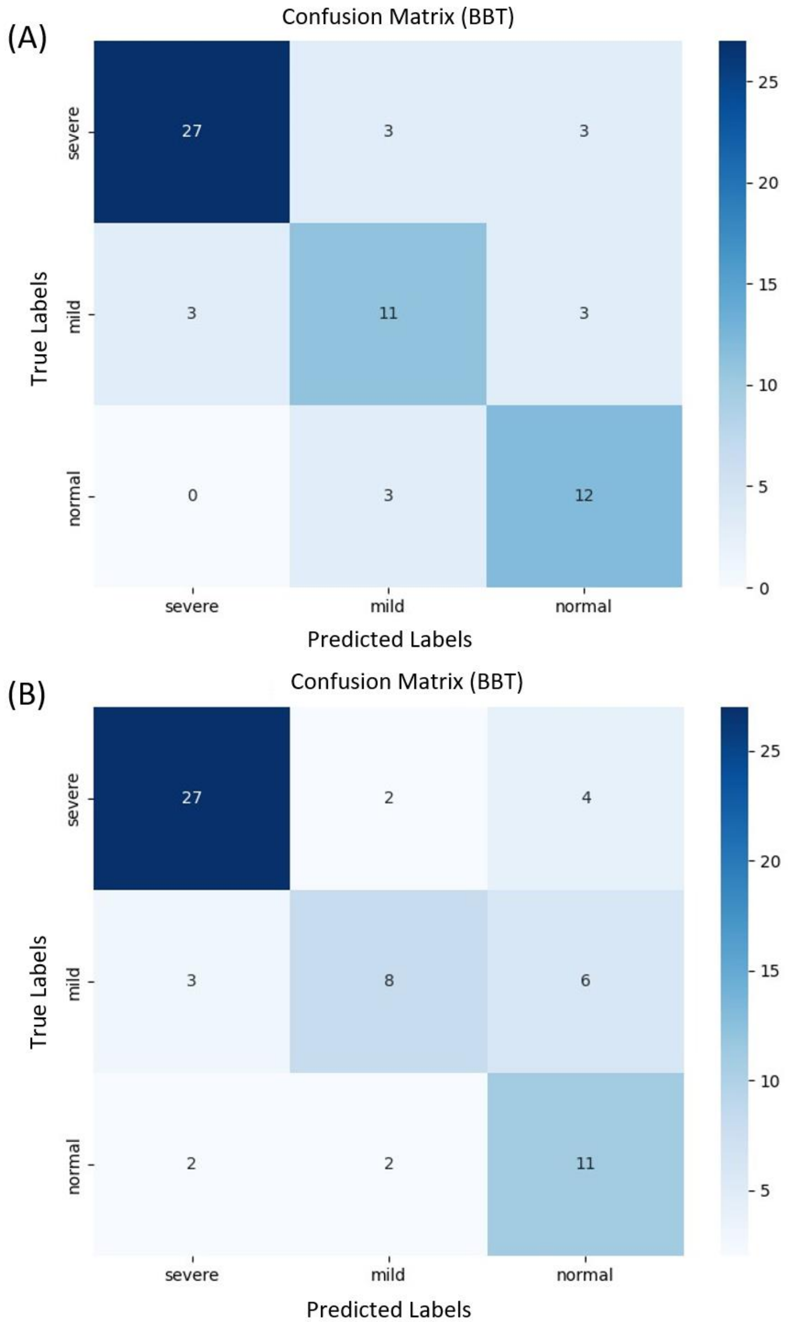


Figure 28. The confusion matrices of the BBT classification using 2D video dataset. (A) CNN, (B) TCN.

C. APS regression

APS was predicted for the 2D keypoint datasets by regression using the CNN and TCN models. Table 7 shows the performance of APS regression using the 2D video data. The correlation coefficient between the APS values predicted for the 2D keypoint datasets and the ground-truth values obtained from 3D motion capture was 0.569 for the CNN model and 0.516 for the TCN model (Figure 29). The mean error between the predicted and ground-truth values was 1.13 for the CNN model and –0.81 for the TCN model (Figure 30).

D. Temporospatial parameters regression

We used the CNN and TCN models to predict movement time, index of curvature, and number of movement units for the 2D keypoint datasets. Table 7 shows the performance of temporospatial parameters regression using the 2D video data. The correlation coefficients between the predicted values of movement time, index of curvature, and number of movement units and the ground-truth obtained from 3D motion capture were 0.528, 0.703, and 0.625 for the CNN model and 0.487, 0.440, and 0.345 for the TCN model (Figure 29). The mean errors between the predicted and ground-truth values of movement time, index of curvature, and number of movement units were 0.21, –0.29, and 1.66 for the CNN model and 0.48,–0.04, and 1.13 for the TCN model, respectively (Figure 30).

Table 7. Performance of regression using 2D video dataset

	Model	Mean error [95% confidence interval]	Correlation coefficient
Arm Profile Score	CNN	1.13 [-8.60, 10.87]	0.569*
	TCN	-0.81 [-11.00, 9.37]	0.516*
Movement times	CNN	0.21 [-3.48, 3.89]	0.528*
	TCN	0.48 [-3.44, 4.41]	0.487*
Index of curvature	CNN	-0.29 [-1.13, 0.56]	0.703*
	TCN	-0.04 [-0.73, 0.64]	0.440*
Number of movement units	CNN	1.66 [-14.79, 18.11]	0.625*
	TCN	1.13 [-14.63, 16.88]	0.345*

[1] CNN, Convolutional Neural Network; TCN, Temporal Convolutional Network

* $p < 0.01$ by Pearson correlation test

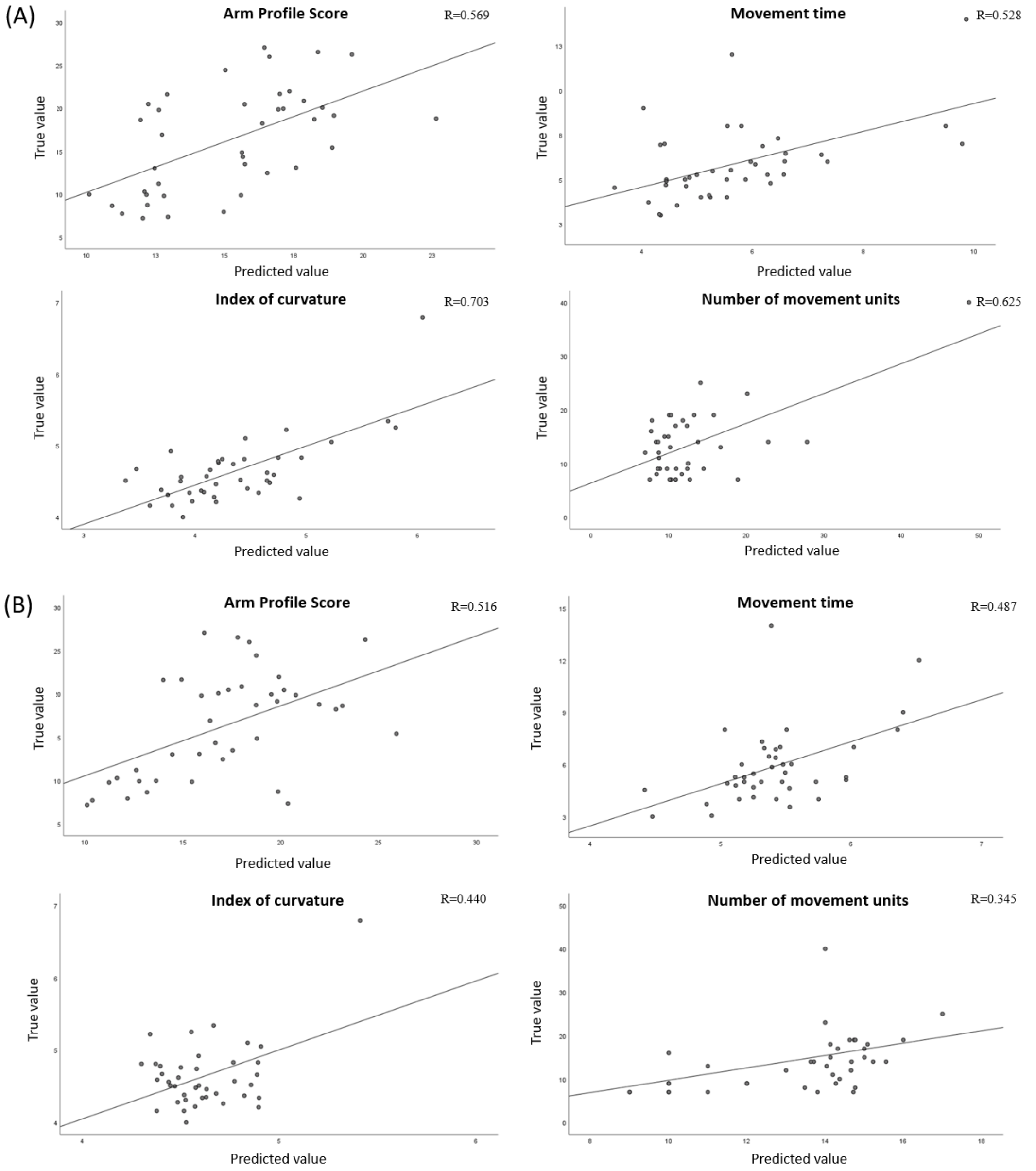


Figure 29. Scatter plots between true values and predicted values by regression using 2D video dataset. (A) CNN, (B) TCN.

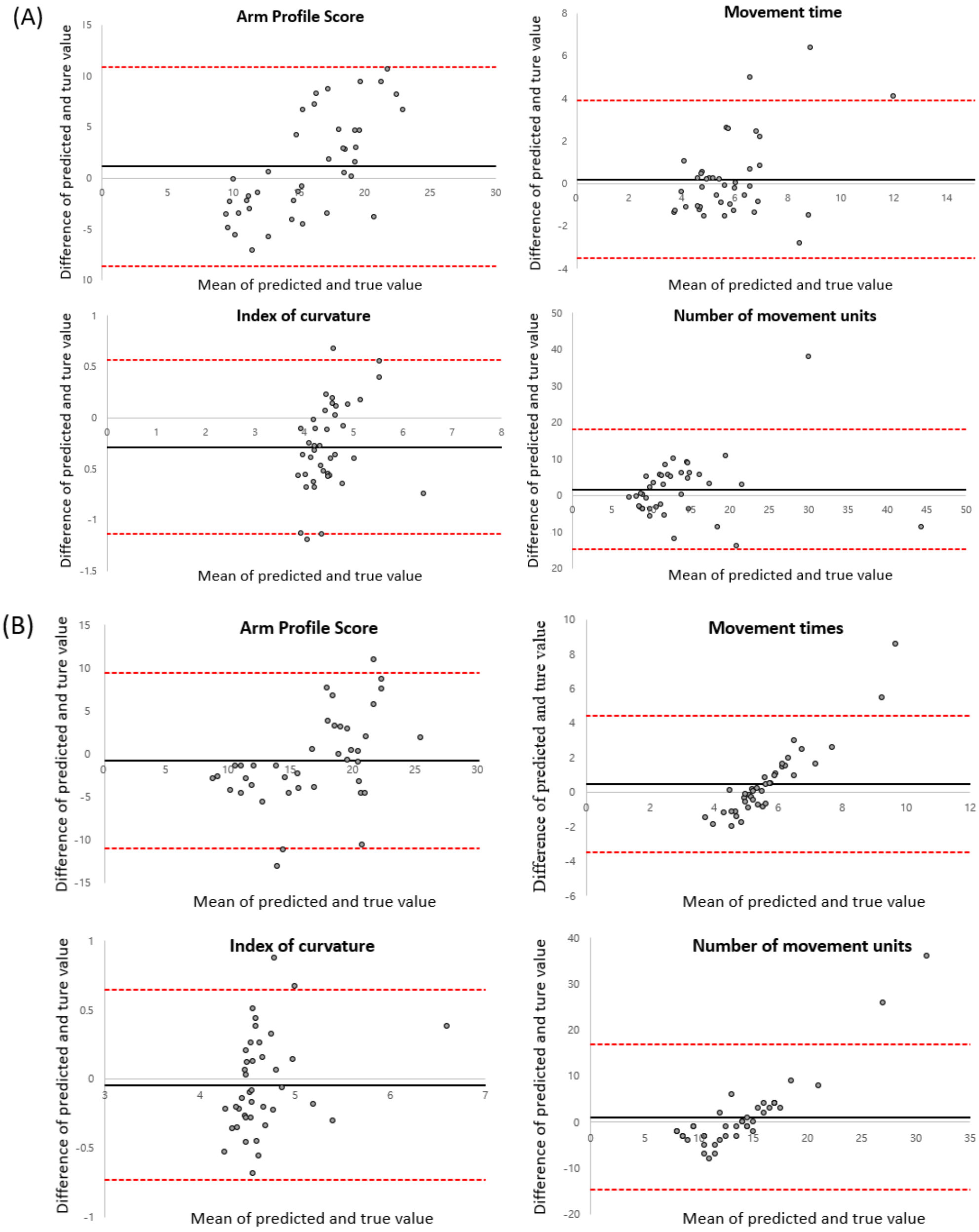


Figure 30. Bland-Altman plots between true values and predicted values by regression using 2D video dataset. (A) CNN, (B) TCN. Black solid line: mean error, Red dotted line: 95% confidence interval.

IV. DISCUSSION

As chronic diseases become more common worldwide as a result of rapid population aging and dietary changes, their social and economic burden is predicted to increase day by day.[36] To counter this increasing burden, digital healthcare based on information and communications technology is emerging as an essential component for future medical systems.[37] Therefore, digital healthcare incorporating technologies such as big data and AI is expected to develop rapidly in the fields of prevention and health monitoring.[21-22] Recently, in the field of health monitoring, studies have been conducted to recognize, classify, and estimate human movements and function by applying deep learning methods to images of human movements.[38] For example, one study used deep learning with gait videos of children with cerebral palsy to successfully predict metrics such as gait speed and gait cadence.[23] Another study used a deep learning model with videos recorded by a depth camera to predict FMA-UE scores in stroke patients with accuracy ranging from 65% to 87%.[39] However, that study required a special camera and used data from only 41 patients, and the FMA-UE estimation by deep learning took the same amount of time as clinical FMA-UE measurement. We aimed to use deep learning methods to contribute to health monitoring in digital healthcare by predicting upper limb function in a relatively large number of stroke patients without space-time constraints and high cost burden.

To predict upper limb function using deep learning, it was necessary to determine parameters that could be used as biomarkers. For this, we used a combination of clinical metrics (FMA-UE and BBT scores) and metrics used in 3D motion capture analysis (temporospatial parameters and APS). The FMA-UE and BBT are tools that have been verified to measure upper limb function in many validation and reliability studies.[4-6] These tools are currently used widely in clinical practice to measure function in patients with stroke. The FMA-UE estimates abnormality of upper limb movement by measuring the range of motion, coordination, and speed of the shoulder, elbow, wrist, and hand. And the BBT is a task-oriented assessment tool that measures limitation of manual dexterity. We attempted to predict not only the overall structure and function of the upper limb but also the ability to perform activity. The Reach & Grasp Cycle performed during 3D motion capture has been verified in many studies as a simple task that reflects the upper limb function of stroke patients.[29-31] In addition, the validity of temporospatial parameters and APS determined by 3D motion capture analysis was confirmed based on correlations with upper limb function scores in children with cerebral palsy.[10-11] Temporospatial parameters were also reported to differ significantly among stroke patients classified as having severe, moderate, or mild upper limb dysfunction based on FMA-UE scores.[26] Therefore, we used 3D motion capture metrics as objective and quantitative biomarkers of upper limb function in our study.

We trained CNN and TCN models to predict upper limb function using time series data measured while patients performed a simple task. CNN models are suitable for learning based on image or time series data, and we used a 1D CNN architecture rather than a 2D CNN architecture for training with time series data.[40] In a previous study, TCN was reported to have better performance than Long Short-Term Memory (LSTM), one of the recurrent neural network techniques suitable for training with time series data.[41] In our study, the CNN model had better overall performance than the TCN model for predicting upper limb function based on time series data of upper limb key points in patients with stroke.

The final goal of this study was to predict upper limb function in stroke patients using deep learning with 2D video data of simple movements recorded by a single camera without expensive equipment. To achieve this goal, we first trained CNN and TCN to predict upper limb function using 3D motion capture datasets.

When FMA-UE classification was performed with a CNN model using 3D coordinate data, the accuracy was 90%. The number of 3D motion capture datasets classified in each FMA-UE category was 234 for severe, 138 for moderate, and 252 for mild. When BBT classification was performed with the same model and data, the accuracy was around 79%. The BBT scores were divided into after normalization by gender and age. The number of datasets classified in each BBT category was 343 for severe, 121 for mild, and 157 for normal. There was no “normal” group for the FMA-UE classification, whereas the BBT classification involved a normal group, so the results suggest that it was difficult to distinguish between mild deficits and normal function using data from a simple task. Also, the confusion matrices of the FMA-UE and BBT classifications indicate that most cases of misclassification involved misclassification into the middle group. These results can be interpreted as classification in the "gray zone," a term commonly used in statistics and classification problems to refer to the middle or border area when trying to classify something into two or more categories or groups.[42-43] This area typically corresponds to cases in which it is difficult or ambiguous to make a clear decision in the classification model.

When APS was estimated with the CNN model using the 3D coordinate dataset, the correlation coefficient between the predicted and ground-truth values was 0.783, indicating a strong correlation. When the movement time, index of curvature, and number of movement units were estimated using the same model and dataset, the correlation coefficients between the predicted and ground-truth values were 0.544, 0.755, and 0.598, respectively, indicating moderate to strong correlations. There was also a strong correlation between index of curvature values predicted by CNN models using the 3D motion capture data and the 2D video data, respectively, with a correlation coefficient over 0.7. Because the index of curvature is calculated using the trajectory of the wrist marker, and the input data are the 3D coordinate values of the marker's trajectory, it can be concluded that the index of curvature, which is highly related to the input data, performed better than the other spatiotemporal parameters.

Our results confirmed the possibility of predicting upper limb function in stroke patients using 3D motion capture data. Therefore, we attempted to predict the upper limb function of the stroke patients using deep learning models with similar features extracted from 2D video data. To implement this, it was important to accurately extract the coordinates of upper limb keypoints from the 2D video. For this purpose, we used a highly accurate pose estimation algorithm. There are various algorithms for estimating human poses in 2D images and videos, such as OpenPose,[44] RTMPose,[33] and HRNet.[45] RTMPose is effective for estimating multiple human poses in 2D video in real time. Pose estimation methods can be broadly divided into two paradigms: top-down and bottom-up. OpenPose follows the bottom-up paradigm by first detecting individual joints of an object in an image and then connecting each joint to estimate the overall pose. Conversely, RTMPose follows the top-down paradigm by first detecting objects and then estimating the position of each joint, making it more accurate than bottom-up methods. A previous study calculated the Average Precision (AP) of several pose estimation methods trained on the COCO-WholeBody dataset.[33] In that study, estimated keypoints were considered correct when they included the actual keypoints within a calculated threshold, object keypoint similarity, and the AP was calculated as the ratio of correct keypoints to estimated keypoints. The AP of RTMPose was 71.2 for body, 57.9 for hand, and 64.8 for whole body, indicating higher accuracy compared with other algorithms such as OpenPose and HRNet.

When FMA-UE classification was performed with the CNN model using 2D keypoint data, the accuracy was 80%. The number of 2D video datasets classified in each FMA-UE group was 138 for severe, 57 for moderate, and 135 for mild. When BBT classification was performed with the same model and data, the accuracy was over 76%, with 172, 87, and 71 datasets classified in the severe, mild, and normal categories, respectively. These results indicate that the performance of the CNN model to classify clinical metrics was lower with the 2D video data than with the 3D motion capture data. However, because the number of data was smaller in the 2D dataset than in the 3D dataset, better performance might be achieved if more data are collected.

When APS was estimated with the CNN model using 2D keypoint data, the correlation coefficient between the predicted and ground-truth values was 0.569, indicating a moderate correlation. The weaker correlation for the 2D data compared with the 3D data might be due to the smaller size of the 2D data. In addition, only the 2D coordinates of each keypoint were estimated. Because the 2D images contain X and Y coordinate information but no depth information corresponding to the Z coordinate, it is difficult to reflect information about the rotation of the wrist or hand. Therefore, there might have been insufficient information to estimate APS, which represents the average value of the RMSE of movement in each plane, including rotation at each joint of upper limb. When the movement time, index of curvature, and number of movement units were determined by regression using the 2D video dataset, the correlation coefficients between predicted and ground-truth were 0.528, 0.703, and 0.625, respectively, indicating moderate to strong correlations. In a previous study, when a CNN model trained with keypoint coordinate data from 2D video recorded in the sagittal plane was used to predict gait metrics, the correlation coefficients for gait speed, gait cadence, and gait asymmetry were 0.73, 0.79, and 0.43, respectively.[23] Gait is a cyclic movement performed repeatedly, and there is little variation in gait motion data among individuals or trials. However, because the Reach & Grasp Cycle is not a continuously repeated task, the variation is greater among individuals and trials. Therefore, the regression performance in our study was better than expected. In addition, the previous study on gait metrics trained the CNN model with more than 1000 datasets. Although our study only used 196 datasets for regression, we observed moderate to strong correlations between the predicted and ground-truth values. If more datasets for regression are added, the performance would be expected to increase further.

We proposed a deep learning method to predict upper limb function in patients with stroke using 2D video of simple upper limb movements. Our method has the advantage of not requiring expensive equipment or trained experts, in contrast to other methods that require specialized instruments or trained clinicians. Our method has the advantage of not requiring expensive equipment or trained experts, in contrast to other methods that require specialized instruments or clinical evaluation. Therefore, our method can enable convenient health monitoring anytime and anywhere using video data of simple movements recorded at home or in daily life using a general camera without trained experts or expensive equipment.

This study has several limitations that need to be addressed before our method can be commercialized for widespread use. First, a small dataset was used to predict upper limb function in patients with stroke. Although it is not easy to collect functional data for patients, our study included a diverse sample of stroke patients with various levels of upper limb function. If data from more patients are collected in future studies, better model performance can be expected. Second, the videos used in this study were taken under controlled conditions in which the angle and position of the camera and the distance between the patient and the camera were fixed. Additional training of the model will be required in cases where the camera settings are different than those used in this study. It is also necessary to verify the model using data recorded at home or in small spaces other than a hospital. Third, our model predicted upper limb function using only spatial coordinate data. To estimate functions other than those predicted in our study, such as muscle tone, muscle strength, and ADL, deep learning models must be developed using data related to those metrics, such as EMG data and other real-world data collected during daily life. Furthermore, data must be collected from patients with stroke over long periods of time to develop a model that predicts not only current function but also future function.

If further studies address these limitations and verify the feasibility of real-world use, a cost-effective measure outside the hospital can be used to evaluate overall upper limb function in patients with stroke, complementing clinical evaluations in hospitals. Additionally, clinicians will be able to remotely track patients' function.

V. CONCLUSION

Several metrics of upper limb function in patients with stroke were accurately predicted by a CNN model using simple 2D video data. Our results suggest that deep learning models can predicted accurately upper limb function measured by time-consuming functional evaluations that require trained clinician or expensive equipment using short and simple movement data. In addition, we showed that deep learning models trained with data from 2D video recorded by a simple camera can achieve performance similar to that of models trained with 3D data recorded with expensive motion capture equipment. These findings provide a basis for further research on the use of digital healthcare to measure and monitor upper limb function in patients with stroke, free from the constraints of time and place.

REFERENCES

1. Fan J, Li X, Yu X, Liu Z, Jiang Y, Fang Y, et al. Global burden, risk factor analysis, and prediction study of ischemic stroke, 1990–2030. Neurology 2023;101 (Pt 2): e137-e150.
2. Duncan PW, Stroke Disability, Physical Therapy 1994;74 (Pt 5): 399–407.
3. Bailey RR, Klaesner JW, Lang CE. Quantifying Real-World Upper-Limb Activity in Nondisabled Adults and Adults With Chronic Stroke. Neurorehabilitation and Neural Repair 2015;29 (Pt 10):969-978.
4. Sanford J, Moreland J, Swanson LR, Stratford PW, Gowland C, Reliability of the Fugl-Meyer Assessment for Testing Motor Performance in Patients Following Stroke, Physical Therapy 1993;73 (Pt 7): 447–454.
5. Allgöwer K, Hermsdörfer J, Fine motor skills predict performance in the Jebsen Taylor Hand Function Test after stroke, Clinical Neurophysiology 2017;128 (Pt 10):1858-1871
6. Platz T, Pinkowski C, van Wijck F, Kim IH, Di Bella P, Johnson G. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. Clinical Rehabilitation 2005;19 (Pt 4):404-411
7. Sears ED, Chung KC. Validity and responsiveness of the jebsen–taylor hand function test. The Journal of hand surgery 2010;5 (Pt 1):30-37.
8. Mackey AH, Walt SE, Lobb GA, Stott NS. Reliability of upper and lower limb three-dimensional kinematics in children with hemiplegia. Gait & posture 2005;22 (Pt 1):1-9.
9. Aggarwal JK, Cai Q. Human motion analysis: A review. Computer vision and image understanding 1999;73 (Pt 3):428-440.
10. Shim D, Choi JY, Yi SH, Park ES, Kim S, Rha DW et al. Spatiotemporal parameters from instrumented motion analysis represent clinical measurement of upper limb function in children with cerebral palsy. Gait & Posture 2022;91:326-

31.

11. Jaspers E, Feys H, Bruyninckx H, Klingels K, Molenaers G, Desloovere K. The Arm Profile Score: A new summary index to assess upper limb movement pathology. Gait & posture 2011;34 (Pt 2):227-33.
12. O'Leary DE. Artificial intelligence and big data. IEEE intelligent systems 2013;28 (Pt 2):96-9.
13. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69, S36-40.
14. Ristevski B, Chen M. Big data analytics in medicine and healthcare. Journal of integrative bioinformatics 2018;15 (Pt 3):20170030.
15. Haleem A, Javaid M, Khan IH. Current status and applications of Artificial Intelligence (AI) in medical field: An overview. Current Medicine Research and Practice 2019;9 (Pt 6):231-7.
16. Yacoub B, Varga-Szemes A, Schoepf UJ, Kabakus IM, Baruah D, Emrich T. Impact of artificial intelligence assistance on chest CT interpretation times: a prospective randomized study. American Journal of Roentgenology 2022;219 (Pt 5):743-51.
17. Kaku A, Parnandi A, Venkatesan A, Pandit N, Schambra H, Fernandez-Granda C. Towards data-driven stroke rehabilitation via wearable sensors and deep learning. Proceedings in Machine Learning for Healthcare Conference; 2020 Aug 6-8; New York, United States: PMLR: 143-71.
18. Alabyad D, Lemuel-Clarke M, Antwan M, Henriquez L, Belagaje S, Nahab F, et al. Telemedicine Impact on Post-Stroke Outpatient Follow-up in an Academic Healthcare Network during the COVID-19 Pandemic. Journal of Stroke and Cerebrovascular Diseases 2023;107213.
19. Tu CC, Weng SY, Hsieh NC, Cheng WC, Alizargar J, Chang KS. Increasing Use of Telemedicine for Neurological Disorders During the COVID-19 Pandemic: A Mini-Review. Journal of Multidisciplinary Healthcare 2023;411-8.

20. Jabarulla MY, Lee HN. A blockchain and artificial intelligence-based, patient-centric healthcare system for combating the COVID-19 pandemic. Healthcare 2021;9 (Pt 8):1019.

21. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International Journal of Environmental Research and Public Health 2021;18 (Pt 1):271.

22. Cirillo D, Valencia A. Big data analytics for personalized medicine. Current opinion in biotechnology 2019;58:161-167.

23. Wang P, Li W, Ogunbona P, Wan J, Escalera S. RGB-D-based human motion recognition with deep learning: A survey. Computer vision and image understanding 2018;171:118-39.

24. Mündermann L, Corazza S, Andriacchi TP. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. Journal of neuroengineering and rehabilitation 2006; 3 (Pt 1):1-11.

25. Kidziński Ł, Yang B, Hicks JL, Rajagopal A, Delp SL, Schwartz MH. Deep neural networks enable quantitative movement analysis using single-camera videos. Nature communications 2020;11 (Pt 1):4054.

26. Choi H, Park D, Rha DW, Nam HS, Jo YJ, Kim DY. Kinematic analysis of movement patterns during a reach-and-grasp task in stroke patients. Frontiers in Neurology 2023;14.

27. Pang MY, Harris JE, Eng JJ. A community-based upper-extremity group exercise program improves motor function and performance of functional activities in chronic stroke: a randomized controlled trial. Arch Phys Med Rehabil 2006; 87:1–9.

28. Mathiowetz V, Volland, G., Kashman, N., & Weber, K. Adult norms for the Box and Block Test of manual dexterity. The American journal of occupational therapy 1985;39 (Pt 6):386-91.

29. Subramanian SK, Yamanaka J, Chilingaryan G, Levin MF. Validity of movement pattern kinematics as measures of arm motor impairment poststroke. Stroke 2010;41 (Pt 10):2303-8.

30. Murphy MA, Willén C, Sunnerhagen KS. Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass. Neurorehabilitation and neural repair 2011;25 (Pt 1):71-80.

31. Rohafza M, Fluet GG, Qiu Q, Adamovich S. Correlation of reaching and grasping kinematics and clinical measures of upper extremity function in persons with stroke related hemiplegia. Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2014 Aug 26-30; Chicago, United States: 2014: 3610-13.

32. Jin S, Xu L, Xu J, Wang C, Liu W, Luo P, et al. Whole-body human pose estimation in the wild.: Proceedings of the 16th European Conference in Computer Vision; 2020 Aug 23-28; Glasgow, United Kingdom: 2020: 196-214.

33. Effective whole-body pose estimation with two-stages distillation. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2-6; Paris, France: 2023: 4210-20

34. Lyu C, Zhang W, Huang H, Zhou Y, Wang Y, Chen K. Rtmdet: An empirical study of designing real-time object detectors. arXiv preprint arXiv 2022:2212.07784.

35. Li Y, Yang S, Liu P, Zhang S, Wang Y, Xia, ST. Simcc: A simple coordinate classification perspective for human pose estimation. Proceedings of the 17th European Conference on Computer Vision; 2022, Oct 23-27; Tel Aviv, Israel: Springer Nature Switzerland: 2022: 89-106.

36. Picco L, Achilla E, Abdin E, Chong SA, Vaingankar A, Subramaniam M, et al. Economic burden of multimorbidity among older adults: impact on healthcare and societal costs. BMC health services research 2016;16:1-12.

37. Haluza D, Jungwirth D. ICT and the future of healthcare: aspects of pervasive

health monitoring. Informatics for Health and Social care 2018;43 (Pt 1):1-11.

38. Sun L, Shang Z, Xia Y, Bhowmick S, Nagarajaiah S. Review of bridge structural health monitoring aided by big data and artificial intelligence: From condition assessment to damage detection. Journal of Structural Engineering 2020;146 (Pt 5):04020073.

39. Kim WS, Cho S, Baek D, Bang H, Paik NJ. Upper extremity functional evaluation by Fugl-Meyer assessment scoring using depth-sensing camera in hemiplegic stroke patients. PloS one 2016;*11*: 7

40. Chen H, Ye W. Classification of human activity based on radar signal using 1-D convolutional neural network. IEEE Geoscience and Remote Sensing Letters 2019;17 (Pt 7):1178-1182.

41. Gopali S, Abri F, Siami-Namini S, Namin AS. A comparison of tcn and lstm models in detecting anomalies in time series data. Proceedings of the IEEE International Conference on Big Data; 2021 Dec 15-18; Sorrento, Italy: 2021: 2415-2420.

42. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. Journal of clinical epidemiology 1990;43 (Pt 1):109-113.

43. Chandra A, Khullar D, Lee TH. Addressing the challenge of gray-zone medicine. The New England Journal of Medicine 2015;372 (Pt 3):203-5.

44. Nakano N, Sakura T, Ueda K, Omura L, Kimura A, Yoshioka S. Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. Frontiers in sports and active living 2020;2:50.

45. Huang J, Zhu Z, Huang, G. Multi-stage HRNet: Multiple stage high-resolution network for human pose estimation. arXiv preprint arXiv 2019:1910.05901.

APPENDICES

Acronyms and Abbreviations

FMA-UE	Fugl-Meyer Assessment-Upper Extremity
BBT	Box and Block Test
3D	Three-dimensional
2D	Two-dimensional
1D	One-dimensional
AI	Artificial Intelligence
APS	Arm Profile Score
SD	Standard Deviation
RMSE	Root Mean Square Error
RTMPose	Real-Time Models for Pose estimation
CNN	Convolutional Neural Network
TCN	Temporal Convolutional Network
leakyRelu	leaky rectified linear unit
MSE	Mean Square Error
LSTM	Long Short-Term Memory
AP	Average Precision

ABSTRACT (IN KOREAN)

뇌졸중 환자에서 간단한 일상생활 동작 데이터로부터 딥러닝에 의한 상지 기능 예측

<지도교수 나 동 욱 >

연세대학교 대학원 의학과

심 다 인

뇌졸중에서 상지 기능을 측정하기 위해 현재 가장 널리 사용되는 방법은 푸글마이어 상지 검사 (Fugl-Meyer Assessment-Upper Extremity; FMA-UE) 및 박스앤 블럭 검사 (Box and Block Test; BBT) 와 같은 임상적 측정법이다. 그러나 임상적 측정은 숙련된 임상의가 필요하고 시간이 많이 소요된다. 뇌졸중 환자의 상지 기능을 측정하는 객관적인 방법은 3차원 상지 동작 분석 검사가 있다. 그러나 이 방법은 비싼 장비와 검사가 가능한 넓은 공간이 필요하다는 한계가 있다. 이러한 기존 방법들의 한계를 해결하기 위해 딥러닝 방법을 이용하여 간단한 일상 동작에서 상지 기능을 예측해보고자 하였다. 따라서 본 연구의 최종 목적은 뇌졸중 환자의 2차원 비디오를 사용하여 딥러닝 방법으로 상지 기능을 예측하는 것이다. 이를 달성하기 위해 2가지 단계를 수행했다. 먼저 뇌졸중 환자에서 3차원 상지 동작 분석 검사로부터 얻은 모션 캡처 데이터를 사용하여 딥러닝 방법으로 상지 기능을 나타내는 지표들을 예측해서 본 연구의 최종 목표의 가능성을 타진했다. 최종적으로는 뇌졸중 환자에서 2차원 비디오에서 자세 추정 알고리즘을 통해 추출한 2차원 키포인트의 좌표 데이터를 사용하여 상지 기능을 예측했다.

본 연구는 후향적 연구로 2014년부터 2023년까지 신촌 세브란스 재활병원에

내원한 265명의 뇌졸중 환자들의 FMA-UE 점수, BBT점수, 움직임 시간 (Movement time, MT), 곡률 지수 (Index of curvature, IC), 이동 단위 수 (Number of movement units, NMU)를 포함하는 시공간적 매개변수들과 팔 프로파일 점수 (Arm Profile Score, APS)를 수집했다. 또한 2021년부터 2023년까지 105명의 뇌졸중 환자에서 뻗기와 잡기 주기 (Reach & Grasp Cycle) 동안 녹화된 2차원 비디오 데이터를 수집했다. 수집된 데이터를 가지고 두 가지 버전의 입력 데이터를 사용하여 딥러닝 모델을 개발했다. 먼저, 3차원 좌표 데이터를 사용하여 상지 기능을 나타내는 지표들을 예측하기 위한 3차원 모션 캡처 데이터셋을 구성했다. 3차원 모션 동안 몸통 (4), 어깨, 팔꿈치, 손목 (2), 손가락 (2)에 총 10개의 반사마커의 X, Y, Z 좌표 데이터로 구성된 총 30개의 좌표 데이터를 얻었다. 두번째, 상지 기능을 예측하기 위해 330개의 비디오 데이터를 사용하여 2차원 비디오 데이터셋을 구성했다. 실시간 자세 추정 모델 (Real-Time Models for pose estimation, RTMPose) 을 이용한 자세 추정을 통해 2차원 키포인트를 추출했다. 하나의 2차원 비디오에서 몸통, 어깨, 팔꿈치, 손목 (2), 손가락 (2)을 포함하는 상지의 7개 키포인트의 X, Y 좌표로 구성된 총 14개의 좌표 데이터를 얻었다. 각각의 입력 데이터를 가지고 합성곱 신경망 (Convolutional Neural Network, CNN)과 시간적 합성곱 신경망 (Temporal Convolutional Network, TCN)을 사용하여 상지 기능 장애의 심각도에 따라 FMA-UE와 BBT를 3개의 그룹으로 분류하고 시공간적 매개변수와 APS를 추정했다. 모든 데이터셋은 훈련과 검증은 위한 데이터셋은 80%, 모델 테스트를 위한 데이터셋은 별도의 데이터인 20%로 분할되었다.

결과적으로 모든 결과에서 TCN보다는 CNN 모델의 성능이 더 좋았다. 먼저 3D 모션 캡처 데이터를 사용한 CNN 모델 학습 결과는 FMA-UE 분류 정확도, 정밀도, 재현율 및 f1 점수 모두 90을 초과했다 (각각 91.13, 90.27, 90.35, 90.31). BBT 분류 정확도, 정밀도, 재현율 및 f1 점수 모두 72를 초과했다 (각각 79.03,

72.54, 73.96, 73.24). 예측된 APS와 시공간 매개변수인 MT, IC, NMU는 참값과 중간에서 강한 상관관계를 가졌다 (r=0.783, 0.544, 0.755, 0.601).

2D 비디오 데이터를 사용한 CNN 모델 학습 결과, FMA-UE 분류 정확도, 정밀도, 재현율 및 f1 점수 모두 85를 초과했다 (각각 89.23, 88.39, 85.97, 87.16). BBT 분류 정확도, 정밀도, 재현율 및 f1 점수 모두 73을 초과했다 (각각 76.92, 73.79, 75.51, 74.64). 예측된 APS와 시공간적 매개변수인 MT, IC, NMU는 참값과 중간에서 강한 상관관계를 가졌다 (r=0.569, 0.528, 0.703, 0.625).

딥러닝 기법은 일상생활의 단순 활동 중에 녹화된 단일 2D 영상만을 이용하여 뇌졸중 환자의 상지 기능을 예측하는 데 매우 유망한 결과를 얻었다. 데이터 수가 작았음에도 불구하고 딥러닝 기법을 사용하여 간단한 동작 하나로 FMA-UE 점수와 BBT 점수를 꽤 정확하게 분류할 수 있었다. 또한 시공간적 매개변수와 APS의 참값과 예측값 사이에 중간에서 강한 상관관계를 보였다. 본 연구에서 딥러닝 기법을 사용하여 아주 간단한 일상 생활의 동작을 촬영한 데이터만으로도 복잡하게 수행되는 기존의 상지 기능 평가 결과를 비교적 정확하게 예측해내었다. 본 연구를 통해 뇌졸중 환자의 간단한 영상 데이터를 딥러닝으로 활용하여 상지 기능을 예측하여, 향후 디지털 헬스케어의 헬스 모니터링 분야에 활용할 수 있는 가능성을 확인하였다.

핵심되는 말 : 뇌졸중, 상지 기능, 예측, 딥러닝, 일상생활동작