

Article

Deep Learning Approaches to Automated Video Classification of Upper Limb Tension Test

Wansuk Choi ¹ and Seoyoon Heo ^{2,*}¹ Department of Physical Therapy, International University of Korea, Jinju 52833, Korea; y3korea@gmail.com² Department of Occupational Therapy, School of Medical and Health Science, Kyungbuk University, Namyangju-si 12051, Korea

* Correspondence: syheo@kbu.ac.kr; Tel.: +82-31-539-5351

Abstract: The purpose of this study was to classify ULTT videos through transfer learning with pre-trained deep learning models and compare the performance of the models. We conducted transfer learning by combining a pre-trained convolution neural network (CNN) model into a Python-produced deep learning process. Videos were processed on YouTube and 103,116 frames converted from video clips were analyzed. In the modeling implementation, the process of importing the required modules, performing the necessary data preprocessing for training, defining the model, compiling, model creation, and model fit were applied in sequence. Comparative models were Xception, InceptionV3, DenseNet201, NASNetMobile, DenseNet121, VGG16, VGG19, and ResNet101, and fine tuning was performed. They were trained in a high-performance computing environment, and validation and loss were measured as comparative indicators of performance. Relatively low validation loss and high validation accuracy were obtained from Xception, InceptionV3, and DenseNet201 models, which is evaluated as an excellent model compared with other models. On the other hand, from VGG16, VGG19, and ResNet101, relatively high validation loss and low validation accuracy were obtained compared with other models. There was a narrow range of difference between the validation accuracy and the validation loss of the Xception, InceptionV3, and DensNet201 models. This study suggests that training applied with transfer learning can classify ULTT videos, and that there is a difference in performance between models.

Keywords: deep structured learning; supervised machine learning; automated feature extraction; Brachial Plexus Tension Tests; rehabilitation medicine; human action recognition



Citation: Choi, W.; Heo, S. Deep Learning Approaches to Automated Video Classification of Upper Limb Tension Test. *Healthcare* **2021**, *9*, 1579. <https://doi.org/10.3390/healthcare9111579>

Academic Editors: Keun Ho Ryu and Nipon Theera-Umpon

Received: 25 September 2021
Accepted: 15 November 2021
Published: 18 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Whereas research into classifying videos using deep-learning approaches has been inclined to be tentative in rehabilitation medicine fields, recent advances in technologies have accelerated research into analyzing overwhelmed video data. Human action recognition has been expected to achieve a more refined and more scientific educational effect in the environment of the recent academic supply, which is described and consumed in images or motion pictures. Video (including images or motion pictures) data are regarded as a spatiotemporal generalization of image data from a traditional neural network's point of view [1], and all neural network structures for image classification have been naturally extended and discussed to a three-dimensional version beyond two dimensions [2]. The machine learning process is a given for deriving insights or making classifications and predictions. It refers to the way the data fit into a mathematical model [3]. Particularly, machine learning discovers patterns that do not involve human subjective judgment or other possible biases from a large amount of data, having high predictive power [4]. Since the introduction of a video classification method using a dimensional convolutional neural network (CNN) [5,6], a 3D CNN has been applied to large-scale video classification. Interestingly, however, the performance of the 3D CNN was only slightly better than that of the CNN, which classified each frame of a video as a 2D convolution. As a result of this, it was

found that important information in video classification is already contained in individual frames, and that information about “movement” in general is not very helpful in classifying videos [7,8]. Therefore, the researchers tried to combine a CNN with a recurrent neural network (RNN) except for a very short video (about 0.5 s) to improve the performance of video recognition, and they achieved excellent results [9,10].

Looking at the review papers on the various methods and algorithms for image or human behavior recognition, we can see in advance that it is very difficult to focus on combining the functions provided by integration of digital transformation (IDT) to increase precision and reduce background noise, and all of these technologies mentioned. Despite its advantages, there are still many limitations in collecting and processing fully dynamic information [10]. An optical flow method can be used to overcome these shortcomings, but it is still not conducive to sports video movements that are captured from different angles. Deep learning-focused behavior recognition methods can handle large datasets, but increasing the efficiency of the system can be considered as a future scope [11].

The key to combining RNN and CNN [12] into one structure is to make all neurons in a convolutional neural network into cyclic units. In other words, it is an extension of the convolutional neural network itself to the cyclic version. Originally, CNNs are more specific feature maps as they ascend from the lower layer to the higher layer, which have excellent performance to distinguish 2D. The RNN is derived from feedforward neural networks, where connections between nodes can represent temporary dynamic behavior by forming directed graphs along a time sequence. RNN can also handle process variable length sequences of inputs using internal state (memory) [13,14]. Models combined to CNN and RNN including ResNet152V2 [15], MobileNetV2 [16], NASNetMobile [17], among others, MobileNetV2 and ResNet50V2 are known to outperform other models for classifying videos such as sports and human behavior.

Deep learning is a technology that instructs computers to perform tasks similar to those performed naturally by the human brain. Based on the research proposed and analyzed on CNN, RNN, LSTM (Long Short-Term Memory), DBN (Deep Belief Network), and GAN (Generative Adversarial Network), which are networks widely used for behavior recognition tasks in previous studies, this study also utilized frameworks for analyzing human activity and behavior is approached to the therapeutic clinical domain [18].

Due to advances in techniques for classifying such videos, research incorporating deep learning into musculoskeletal fields [19] might be of interest. Given the recent interest in musculoskeletal video data collection [20,21], it is predicted that machine learning or deep learning research using these data would subsequently expand to orthopedic or occupational therapy. Especially due to the coronavirus pandemic, many practical classes have been canceled or reduced, and classical physical training methods are usually the only alternative [22], raising complaints from educators and students; learners are still demanding smart educators who can guide musculoskeletal work in non-face-to-face conditions [23].

Most of the clinical skills and technical parts used in the rehabilitation medicine field are apprenticed, which takes a long time and individual deviations are large because students or interns acquire them from a first-person perspective in terms of education. The image information on the posture of each step shown from the third person point of view may be a timely attempt in the era of a new technology.

In Korea, even in the physical therapy curriculum which currently requires actual practice, face-to-face education is being disrupted due to the spread of the COVID-19 pandemic, and most students rely on superficial video recordings or attend online lectures. In such a circumstance, an accurate image/video analysis model would be able to provide students with learning contents based on accuracy close to actual practice.

Analysis of sports performance during physical activity is an important indicator used to improve the performance of players during a match. In this paper, a review study on video-based technology for sports action recognition for building an automated notation analysis system was introduced and the principle was identified [24]. However, in this

study, it is more meaningful and an essential process for the practical stage because it is not limited to theory, but actually grafted such research into actual clinical practice.

As one of these solutions, we attempted to build a deep learning model that recognizes Upper Limb Tension Test (ULTT) videos, which is used as one of the evaluation tools for manual therapy. In this process, we tried to compare several transfer learning models to determine which model is the best at ULTT video recognition. In this work, we compared the performance with existing models by fine tuning models such as NASNet-Mobile, InceptionV3, and Xception, which were newly released but rarely used for video classification. We aimed to find a good model that can recognize Upper Limb Tension Test (ULTT) behavior well.

2. Methods

2.1. ULTT Clinical Settings

Pain in the neck, shoulders, and upper limbs, which is commonly encountered clinically, is directly related to the nervous systems. Neurological defects or lesions are highly associated with upper arm neurosurgery, and there is an upper limb neurological test (ULTT) designed by Elvey (1980) as a method to clinically identify these signs [25]. In particular, when the adaptive mechanism of the nervous system is damaged, abnormal muscle tension such as contracture, shortening, and myoclonus occurs in the muscle, and joint range of motion is limited, and sensory disturbance occurs. In addition, when peripheral nerve damage occurs, additional disorders such as restriction of muscle activity, edema, blood circulation disorder, and autonomic nervous system dysfunction such as sweating occur concurrently [26]. Pain in the neck, shoulder, and upper extremities that are commonly encountered clinically is directly related to the nervous system. Neurological defects or lesions are highly correlated with the brachial plexus, and the upper limb tension test (ULTT) devised by Elvey (1980) is a clinical method to confirm these signs [27]. This tension test is to create a posture of maximum tension by gradually applying step-by-step tension to various nerve structures up to the maximum range of motion. Upper Limb Tension Test 1 (ULTT1) is test for median nerve, anterior interosseous nerve, and clinician lowers the shoulder of the anterior interosseous nerve and spreads it 110 degrees. The elbows, wrists, and fingers are straight, and the forearms are twisted. Upper Limb Tension Test 2A (ULTT2A) is another type of median nerve bias test, which is a test operation in the order of the musculocutaneous and axillary are lowered and turned lateral to 10 degrees apart. The forearms, wrists, and fingers are straightened, and the forearms prostrate. In this position, the median nerve, musculocutaneous nerve, and axillary nerve are tense. Upper Limb Tension Test 2B (ULTT2B) is a test for radial nerve bias and is a test in order of shoulder girdle depression, elbow extension, medial rotation of the whole arm, wrist, finger, and thumb flexion. Lastly, Upper Limb Tension Test 3 (ULTT3) is a test for ulnar nerve bias, which is a test movement in lower the patients' shoulders and rotate them inward to 10 degrees apart. The elbows are straightened, the forearms are propped up, and the wrists are bent so that they are tilted toward the left side, and the fingers are also bent. In this posture, the old nerve is tense (Figure 1). In the text, representative pictures were presented, and in the actual video, each therapist was examined in a slightly different posture. For example, one clinician pressed the shoulder to increase nerve tension, whereas another instructed the patient to move the neck flexion to the left or right. The specific implementation procedures of ULTTs performed by clinicians were summarized using existing references [25,28] and presented at each stage through Table 1 of this paper.

2.2. Deep Structured Learning Experimental Settings

Deep learning model experiments for video classification and analysis were conducted in Python 3.8, Opencv 4.2, Keras 2.4, and TensorFlow 2.0 environments. In addition, TensorFlow-gpu 2.0 was used to utilize the effective GPU processor in training time. The specifications used in the experiment were Intel® Xeon® Gold 5120 CPU © 2.20 GHz,

180 GB RAM, two NVIDIA Tesla V100-SXM2-32GB, and Windows 10 64 bit operating system was installed.

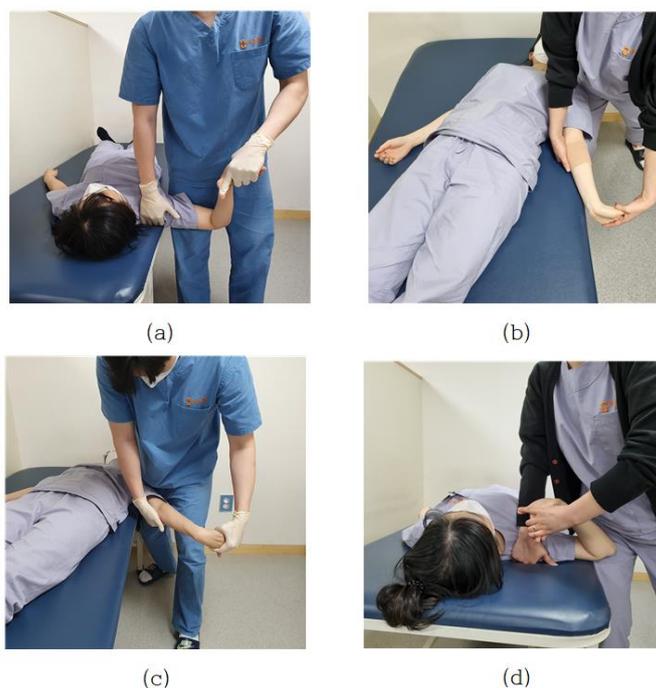


Figure 1. (a) Upper Limb Tension Test 1 (ULTT1); (b) Upper Limb Tension Test 2A (ULTT2A); (c) Upper Limb Tension Test 2B (ULTT2B); (d) Upper Limb Tension Test 3 (ULTT3).

Table 1. Detailed Process of Upper limb tension tests (ULTTs).

ULTT 1	ULTT2A	ULTT2B	ULTT3
<ol style="list-style-type: none"> Starting position. Note patient's thumb and fingertips supported, plus some of the weight of the arm taken on the therapist's thigh. Shoulder abduction to symptom onset, or tissue tightness, or approximately 100 degrees. Wrist extension. Make sure the shoulder position is kept stable. Wrist supination, again making sure that the shoulder position is kept stable. Shoulder lateral rotation, to symptom onset or where the tissues tighten a little. Elbow extension to symptom onset. Neck lateral flexion away, making sure it is whole neck and not just the upper cervical spine. Neck lateral flexion towards. This should ease evoked symptoms. 	<ol style="list-style-type: none"> Patient has her shoulder girdle just over the side of the bed Shoulder girdle depression (via the therapist's thigh) to symptoms or where the tissues tighten a little Elbow extension Whole arm lateral rotation, keeping shoulder girdle depressed Wrist and finger extension (note suggest grip in the inset) 	<ol style="list-style-type: none"> The patient lies with their shoulder just over the side of the bed, the therapist uses his thigh to carefully depress the shoulder girdle Elbow extension Notice how the therapist has brought his left arm 'around' to grasp the patient's wrist in order to medially rotate the whole arm Whole arm medial (internal) rotation Wrist and thumb flexion can be added. Leave the fingers out as the extensors will be too tight Adding a few degrees of shoulder abduction will sensitize the test and elevation of shoulder girdle will provide structural differentiation 	<ol style="list-style-type: none"> Starting position—let patient's elbow rests on the therapist's hip Wrist and finger extension, ensure 4th and 5th fingers are extended Pronation Shoulder lateral rotation, ensuring wrist position is maintain Elbow flexion Block shoulder girdle elevation by pushing first into the table

2.3. Video Collection

Two independent external physical therapists who are not included in the study authors carefully selected only the correct data from four type of ULTT videos (ULTT1, ULTT2A, ULTT2B, and ULTT3) on YouTube. Search terms are upper limb tension test, ULTT, upper limb tension test 1, ULTT1, median nerve bias, upper limb tension test 2A, ULTT2A, median nerve bias, upper Limb tension test 2B, ULTT2B, radial nerve bias, upper limb tension test 3, ULTT3, ulnar nerve bias. Each of the four classes consisted of 130 video clips. A Python program for collecting video address, 4k Video Downloader, Fast Duplicate File Finder, Free Video to JPG Converter and Windows photo editing program were used in the video image download, editing, and frame extraction process.

2.4. Dataset and Preprocessing

An image was extracted every frame from this video, and each image was resized to (224, 224, 3) for useful batch learning. In addition, the data augmentation technique was applied to reduce generalization loss and to make the model more powerful. Each class of the dataset is ULTT1, ULTT2A, ULTT2B, and ULTT3, and consists of 131,458 images. The process of moving and examining joints in the ULTT image frame was analyzed, and frames including actions and postures unnecessary for training were removed. In the entire ULTT images dataset of 103,116 frames, the ratio of training and validation frames was assigned to 8:2 (Figures 2 and 3).

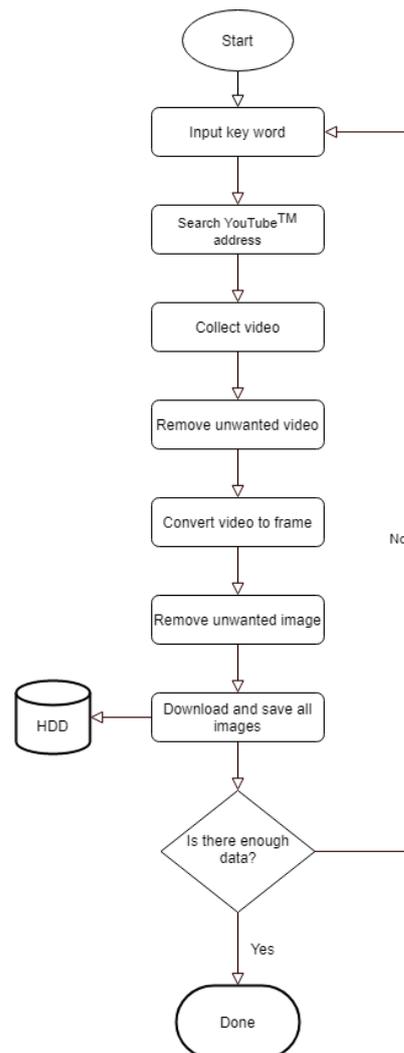


Figure 2. Image dataset collection process.

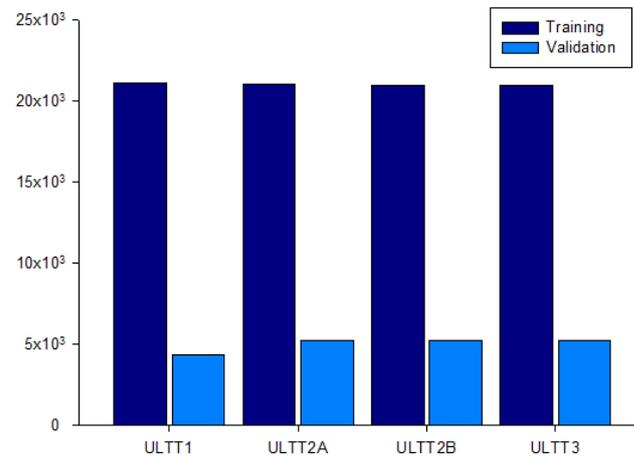


Figure 3. Dataset for the distribution of 103,116 ULTT images.

2.5. Working with the Dataset

2.5.1. Extracting Features from the Frames Using CNN

ULTT video is one of the physiotherapy techniques for examining a patient, and it is difficult to analyze the image in engineering due to the surrounding environment, image quality, camera angle, and various gestures. Therefore, since it is difficult to estimate motion features with existing methods, we tried to estimate accurate ULTT motions using a deep learning-based classification method.

CNN (Convolutional Neural Network) is a technology that performs two-dimensional image analysis and classification prediction based on an artificial neural network model in the field of computer vision. In this study, we also tested ULTT motion estimation using CNN. Figure 4 shows the fundamental model structure for learning the ULTT dataset, and each fine-tuning model was trained. In the process of the experiment, we tried to achieve the best training result by designing a dense layer suitable for dataset learning through trial and error and setting hyperparameters.

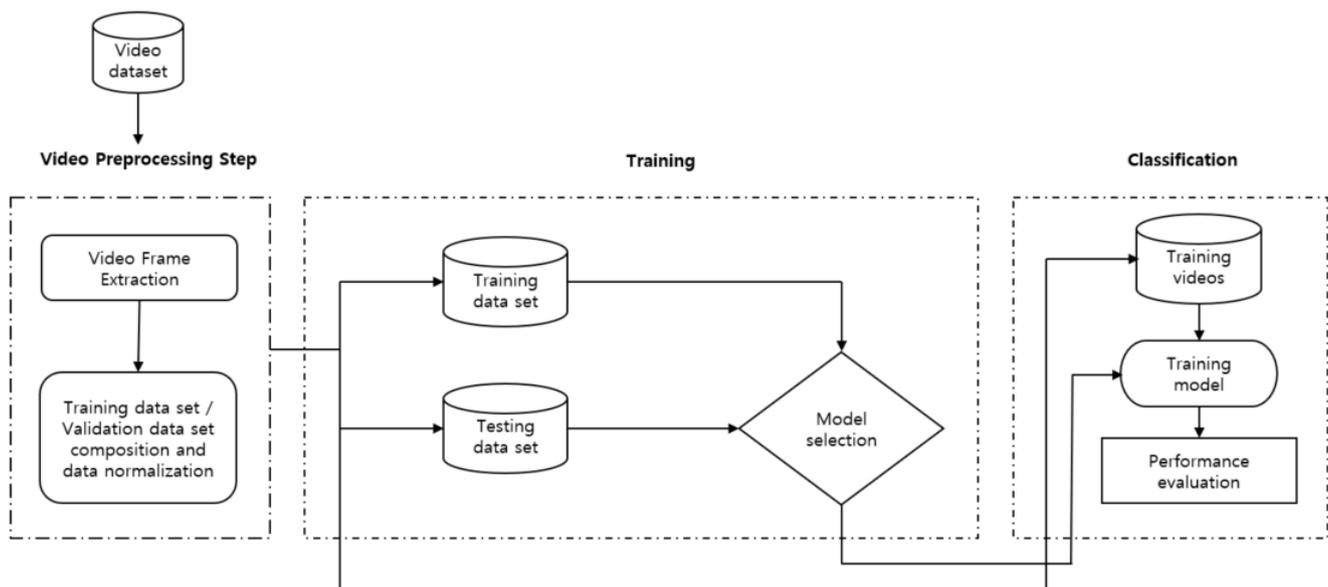


Figure 4. Proposed framework for human action recognition.

The pretrained CNN models DenseNet121, DenseNet201, InceptionV3, NASNetMobile, ResNet101, VGG16, VGG19, and Xception provided by Keras was used to extract feature vectors from the ULTT videos. Using the weights of these models pre-trained on

the ImageNet dataset consisting of 1000 classes (1.2 million labeled images) contributed to saving a lot of time and effort to obtain results (Figures 4 and 5).

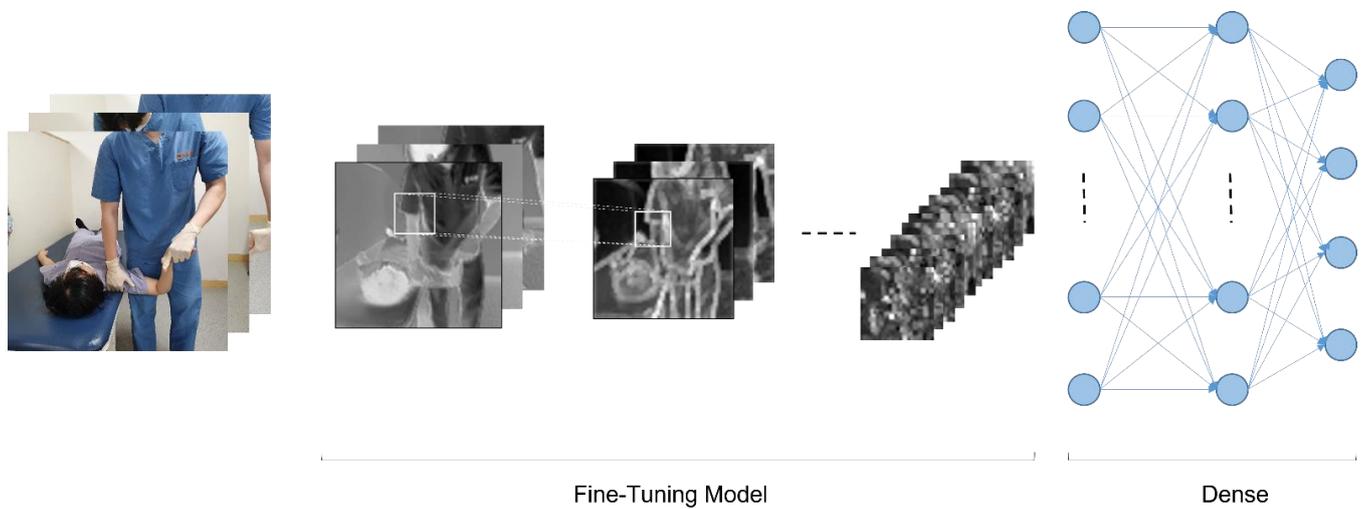


Figure 5. Basic model structure for video classifications.

We derived the classification probability $cost_i$ by adding a density layer at the end, as shown in the following Equations (1) and (2) for probabilistic calculation with 4 types of classes which could lead to minimize loss functions.

$$cost_i = \frac{1}{m} \sum_{i=0}^m (\text{dense}(\text{model}(I_i)) - y_i) \tag{1}$$

$$\text{dense}(I_i) = \text{ReLU} \left(b_j + \sum_{i=0}^n I_i \times w_{ij} \right) \tag{2}$$

Here, I_i is the training image, $\text{model}(I_i)$ is each transfer learning model, and y_i is the predicted classification probability (ground truth or label). In this training, a learning rate of 0.0001, batch size of 32, and epochs of 100 were used by fine-tuning hyperparameters.

2.5.2. Loss Function

The loss function is the ‘difference or error between the predicted value and the correct value’ and was used to optimize the neural network as a function that was minimized by the selected optimizer. In this process, multiple classification is performed by supervised learning, and the correct answer class is expressed as a label value (e.g., 0, 1, 2, 3, . . .) rather than a ‘one-hot encoding’, thus the ‘sparse_categorical_crossentropy’ loss function was used. Additionally, since it is a multiple classification, ‘Softmax’ (Figure 6) was used as the activation function of the output layer (Equation (3)).

$$f(s)_i = \frac{e^{s_i}}{\sum_j^c e^{s_j}} \quad CE = - \sum_i^c t_i \log(f(s)_i) \tag{3}$$

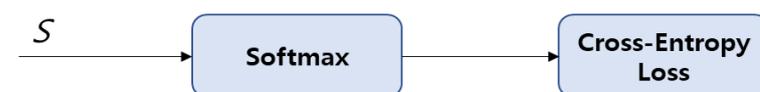


Figure 6. Flow diagram of activating function from ‘Softmax’.

2.5.3. Saving the Best Model and Classifying Videos

The authors saved the model weights for the iteration whenever the validation loss was the smallest, because saving the model only with the best accuracy, regardless of the

loss, would not result in the best model. Therefore, for this study, the smallest loss was used to specify the best model. Then, we defined the same architecture we defined as training to classify the video. Next, we loaded the model weights, and then we were ready to classify the videos. This involved obtaining the video, extracting the frames, and then using the CNN model to extract features from those videos by passing these features to the architecture.

3. Results

The results showed almost similar progress; training loss and validation loss showed a tendency to increase in value from the Xception to the ResNet101, whereas training account and validation account showed a decreasing direction. Validation loss and validation accuracy of Xception, inceptionV3, and DenseNet201, which are evaluated as models with excellent performance, are similar between models, but there is a very small difference in validation loss at three decimal places and validation accuracy at four decimal places. The NASNetMobile and DeseNet121 models showed intermediate performance, and the validation loss of the NASNetMobile model was slightly lower and the validation accuracy was slightly higher. The validation accuracy of VGG19 and ResNet101, which showed low performance, were similar to each other, but the validation loss was significantly lower in VGG19 (Table 2, Figures 6–8).

Table 2. Model Performance comparison for transfer learning from ULTT datasets.

Model	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Xception	0.0012	0.9999	0.0014	0.9999
InceptionV3	0.0016	0.9998	0.0024	0.9996
DenseNet201	0.0037	0.9998	0.0033	0.9996
NASNetMobile	0.0151	0.9977	0.0173	0.9967
DenseNet121	0.0181	0.9972	0.0197	0.9965
VGG16	0.1962	0.9619	0.1973	0.9605
VGG19	0.242	0.9491	0.2418	0.9467
ResNet101	0.6044	0.8093	0.6053	0.8102

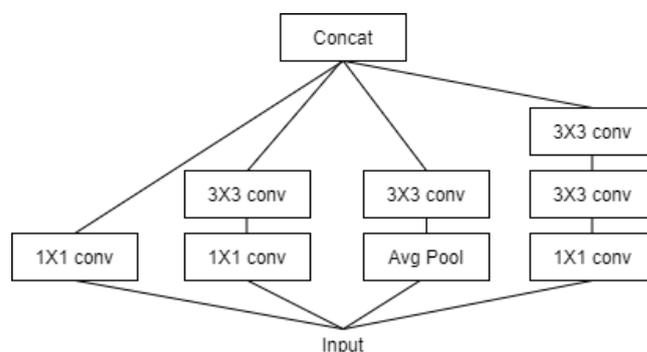


Figure 7. Architecture of Inception V3 Model.

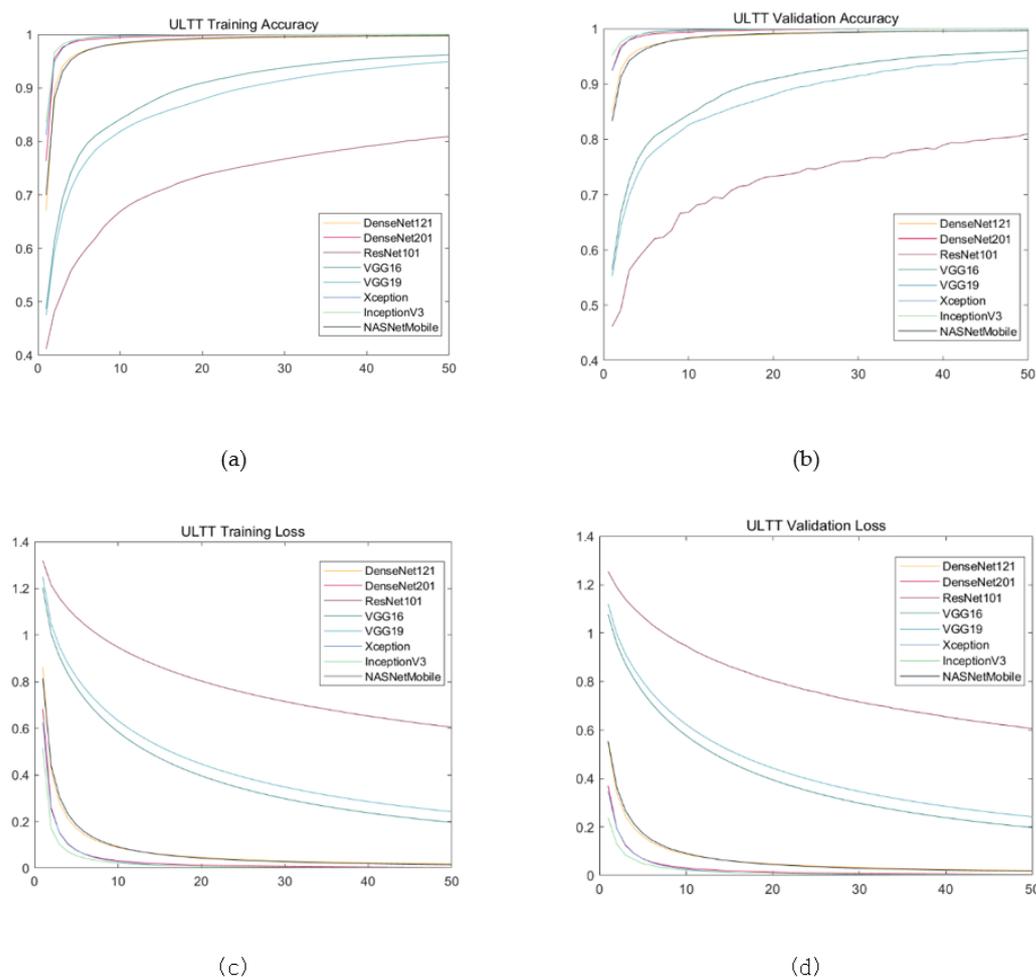


Figure 8. Training results on ULTT data set. (a) Training accuracy; (b) validation accuracy; (c) training loss; (d) validation loss.

4. Discussion

ULTT is an upper limb tension test and a representative evaluation method that can clinically identify nerve muscle shortening and stiffness [29]. In particular, when the adaptation mechanism of the nervous system is damaged, abnormal muscle tension such as contracture, shortening, and myoclonus occurs in the muscle, and limitation of the joint range of motion and sensory disturbance also occur. In addition, when peripheral nerve damage occurs, additional disorders such as restriction of muscle activity, edema, blood circulation disorder, and autonomic nervous system dysfunction such as sweating occur concurrently. Pain in the neck, shoulder, and upper extremities commonly encountered clinically is directly related to the nervous system. Neurological deficits or lesions have many associations with the brachial plexus, and ULTT was adopted as a clinical method for identifying these signs [30]. ULTT, which is evaluated as a simple yet clinical core technique [31], was used to see that it can be implemented in a therapeutic environment beyond the realm of setting the realm of human action recognition in sports or motion itself. These attempts have not been actively attempted in other papers, and are expected to be used to maximize communication between medical staff in education and clinical settings.

In this study, the performance of eight types of models that classify four types of ULTT video through transfer learning was compared. During data training, as the number of epochs increased, training and validation accuracy increased while training and validation loss decreased. For models with good performance such as Xception, InceptionV3, and DensNet201, the accuracy increased sharply between 20–30 epochs and the loss decreased sharply. However, in the case of models such as VGG16, VGG19, and ResNet101, the

degree of increase in accuracy or decrease in loss was relatively modest. Moreover, in the case of the excellent model, validation and training loss became constant after 10 epochs, and validation and training accuracy of 97–99% were obtained.

Compared with this study, the validation loss (error) of the study classifying sports videos using transfer learning was lower in the VGG16 model than in the VGG19 [32]. It would be presumed that these different results were probably due to differences in datasets and parameters. Meanwhile, another study that categorizes various sports (18 types) using the VGG16 transfer learning model had a validation accuracy of 0.93 and a validation loss of less than 0.25, indicating that the validation accuracy was low and the validation error was higher than the results of our study [33]. This suggests that the performance of VGG16 applied to the dataset of this study would be surpassed.

In addition, the validation loss of ResNet101, VGG16, and VGG19 decreased gently as the epochs increased, whereas the validation loss values of other pre-trained models including Xception mostly fell to the shortest before 20 epochs. The validation loss values of ResNet101, VGG16, and VGG19 of these three models were relatively higher than those of other learning models. These results suggest that DensNet121, DensNet201, Xception, InceptionV3, and NASNetMobile models are suitable and sophisticated models for the dataset from this study. In addition, the validation loss of the ResNet101 model was the highest among the models with a gentle fall, and the validation accuracy increased unstable as the process progressed, and the value was also the lowest.

Xception and InceptionV3 showed the best performance among eight models. Xception is further derived from the InceptionV3 model and obtained the highest performance (numerical) among the models. The InceptionV3 model uses a method to simultaneously map each cross-channel correlation and spatial correlation by adding a channel dimension to the convolutional layer as shown in Figure 7. For example, different filter results are obtained at each step for features such as human eyes, nose, mouth, arms, and joints, and classification probability is estimated based on this. Certain types of deep learning models should learn well by increasing the depth of the hidden layer, but in the case of a hidden layer that is too deep, there may be problems such as vanishing/exploding gradients, etc. Therefore, the Xception model developed to solve this problem solved the learning problem by preventing vanishing/exploding gradients by adding the previous identity data through the residual connection method. In the ULTT dataset, information analysis on the direction of the arm and joints is important, and our team obtained interesting and satisfactory results as the Xception model performed very well in classifying and analyzing each piece of delicate feature information through cross-correlation.

Unexpectedly, ResNet101 has performed favorably in studies for human action recognition [34,35]. In this study, ResNet101 model performed significantly lower compared with other research, due to the fact that it is estimated that the noise of background information in the ULTT dataset interferes with its training. In addition, models other than ResNet101 applied in this trial presented advantageous performance on the ULTT dataset; however, when the validation accuracy approaches 1.0 and the validation error approaches 0.0, it would be determined that they could be analyzed as generalization errors.

Deep learning-based solutions for computer vision have made it easier to deliver technical content for educational purposes or clinical use. It has been mentioned in other papers that conventional image data contain many hidden pieces of information and patterns that can be used for human activity recognition (HAR), and HAR can be applied to many areas such as behavioral analysis, intelligent video surveillance, and robot vision [36]. In terms of erroneous classification, existing hand-engineered and machine learning-based solutions have little or no ability to handle overlapping tasks. A fine-tuned pretrained CNN model learns spatial relationships at the frame level [11,18,37]. In this paper, in order to derive an optimized method, the video classification performance was compared with eight transfer learning models (Xception, InceptionV3, DenseNet201, NASNetMobile, DensNet121, VGG16, VGG19, ResNet101) that are most reliable at the current time point. It

is believed that the aforementioned loss could be greatly reduced or corrected by repeating their verification accuracy and loss testing.

Other studies have used so-called HAR, which automatically identifies the physical activity that humans perform, to infer the behavior of one or more people from a series of observations captured by sensors, video, or still images [38]. Recognizing human action in a video sequence is a very difficult task due to issues such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and shape. In this study, to improve these shortcomings, CNN models are actively used to show improved performance compared with some existing deep learning-based models.

Although this study presented approximately satisfactory results, it has several limitations. The labeling of the dataset constructed in this paper is for the purpose of simple classification. Therefore, since the classification of the four types of ULTT data set is estimated as a probability value between 0 and 1, more accurate classification performance was obtained in ResNet101, Dense121, etc., which were used before the recently published model.

Despite many efforts, the test accuracy scores in this study seemed to be not very significant. Our model did not generalize to the test set because it could not accurately predict the first data it presented. Our technical team analyzed several possible causes of lowering test accuracy. First, data leakage was not the cause of the lower test accuracy. The training, validation, and test sets were randomly sampled, and the test set was divided, which may lead to data leakage. However, we decided that data leakage was not the cause of lowering test accuracy, as we aimed to make it similar to the data we had never seen before in the light of motion from humans. Second, it is possible that overfitting had occurred. This is according to the more complex the model, the more accurate it can predict on the training data, but when quite complex it becomes too sensitive to each data point in the training set and does not generalize well to new data. Third, we depended only on pretrained models such as Xception for video classification. If libraries such as OpenPose and Keypoint Detection were used, it is expected that the test accuracy score would be higher by analyzing the joint movement more accurately.

Deep learning technology does not mean better creation of missing information. More precise prognosis could be predicted only when the various factors that affect the patient's prognosis [39] are collected more closely. There would be a limitation to learning sufficient prior knowledge from various input data [40]. A method of giving prior knowledge to a deep learning model has also been proposed, but this is to inform the model with prior knowledge, and ultimately there may still be a problem in which the model cannot be trained from data, and the prior knowledge given in this way is renewed for every problem. There may be issues that need to be defined.

Patients' various clinical information must be highly accurate and data-driven in a variety of ways to enable the realization of true Precision Medicine [41]. Deep learning technologies in this research would not perform efficiently compared with other machine learning techniques as Artificial neural network (ANN), Support Vector Machine (SVM), Logistic Regression, Multi-Layer Perceptron (MLP), and Random Forest.

There is a limitation to the technical transfer that takes place within the institution to which apprentices or novice clinicians belong when they are in the training process or upgrading their skills. It is considered that educational advancement in posture and clinical technology through various images and image analysis could be introduced through this study.

5. Conclusions

The systematic design and modeling of sequential deep learning for human action recognition is expected to be a tool to help with big data processing and educational purposes, which are important in the current academic communication methods described in images. Generally, clinical trials in the medical field using deep learning mechanisms have a fundamental purpose in making accurate predictions of patients' prognosis or in

making judgments for accurate diagnosis [23,26]. In this study, in a similar context, it was designed to provide a scientific basis for technical judgment and clinical decision-making processes that should be applied differently depending on the patient's condition.

Sufficient numbers and quality data are needed for machine learning techniques to show stronger performance than statistical techniques [40], according to the fact that most clinical metadata could be numerous and unrefined. High-quality cohort data may be needed to fully utilize the strengths of deep learning [41].

In this study, the performance of several pretrained models, including Xception, to classify ULTT video was compared. Unfortunately, despite obtaining high validation accuracy and low validation loss scores in most models except for ResNet101, the test accuracy scores may seem insufficient in the degree of completion compared with other similar studies. Although we judged that the dataset has class balance and only presented accuracies, in the next study, precision, recall, and F1 score will be additionally presented so that model performance can be evaluated in more detail. Additionally, as a study to analyze classification problems, we will be able to present a confusion matrix together. Additionally, we will try to find the model at the optimal point where the generalization performance is at maximum. If learning using an advanced library (OpenPose, Keypoint Detection, etc.) and advanced dataset composing key point labels or line labels is attempted, more improved video classification results and image analysis results will be obtained. Furthermore, follow-up research could apply the recently introduced Full Convolution Dense Dilated Network (FCdDN) model, which achieves a favorable segmentation efficiency performance while ensuring high accuracy [42]. We aim to explore more diverse ULTT datasets in order to improve the model's performance. Once this method is applied, it would be possible to promote the convenience of learners and teachers by recognizing various treatment and examination movements.

Author Contributions: W.C. conceived conceptualization, writing original draft preparation, collaboration with technicians and experts; S.H. performed writing original draft preparation, project administration, writing—review and editing and investigation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available from the corresponding author S.H. on request.

Acknowledgments: The authors are indebted to Haneul Yu, a researcher from the multimedia computing laboratory, Chung-Ang University, Korea, for his technical support of the computing process. We would like to thank South Korea's 'National IT Industry Promotion Agency' for lending high-performance computers without any strings attached, 'Busan Industry Promotion Agency, ymt Corp.', Eun-Ha Kong from 'HealthnAI' for professional consulting on deep learning analysis and thesis preparation. A-Jeong Kim and Yu-Sung Jang assisted with the graph and photo-related works.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Taylor, G.W.; Fergus, R.; LeCun, Y.; Bregler, C. Convolutional Learning of Spatio-Temporal Features. In *Computer Vision—ECCV 2010*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 140–153.
2. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
3. Woo, D.-C.; Moon, H.S.; Kwon, S.; Cho, Y. A Deep Learning Application for Automated Feature Extraction in Transaction-based Machine Learning. *J. Inf. Technol. Serv.* **2019**, *18*, 143–159. [[CrossRef](#)]

4. Jordan, M.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260. [[CrossRef](#)] [[PubMed](#)]
5. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential Deep Learning for Human Action Recognition. In *Human Behavior Understanding*; Salah, A.A., Lepri, B., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
6. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [[CrossRef](#)] [[PubMed](#)]
7. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
8. Ng, J.Y.-H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond Short Snippets: Deep Networks for Video Classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
9. Tran, A.; Cheong, L.-F. Two-Stream Flow-Guided Convolutional Attention Networks for Action Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 3110–3119.
10. Baisware, A.; Sayankar, B.; Hood, S. Review on Recent Advances in Human Action Recognition in Video Data. In Proceedings of the 2019 9th International Conference on Emerging Trends in Engineering and Technology—Signal and Information Processing (ICETET-SIP-19), Nagpur, India, 1–2 November 2019; pp. 1–5.
11. Roubleh, A.A.; Khalifa, O.O. *Video Based Human Activities Recognition Using Deep Learning*; AIP Publishing LLC.: Shah Alam, Malaysia, 2020; p. 020023.
12. Ballas, N.; Yao, L.; Pal, C.; Courville, A. Delving Deeper into Convolutional Networks for Learning Video Representations. *arXiv* **2016**, arXiv:1511.06432.
13. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshad, H. State-of-the-Art in Artificial Neural Network Applications: A Survey. *Heliyon* **2018**, *4*, e00938. [[CrossRef](#)] [[PubMed](#)]
14. Tealab, A. Time Series Forecasting Using Artificial Neural Networks Methodologies: A Systematic Review. *Future Comput. Inform. J.* **2018**, *3*, 334–340. [[CrossRef](#)]
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 630–645.
16. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
17. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning Transferable Architectures for Scalable Image Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8697–8710.
18. Pareek, P.; Thakkar, A. A Survey on Video-Based Human Action Recognition: Recent Updates, Datasets, Challenges, and Applications. *Artif. Intell. Rev.* **2021**, *54*, 2259–2322. [[CrossRef](#)]
19. Tack, C. Artificial Intelligence and Machine Learning | Applications in Musculoskeletal Physiotherapy. *Musculoskelet. Sci. Pract.* **2019**, *39*, 164–169. [[CrossRef](#)] [[PubMed](#)]
20. Bedard, N.A.; Schoenfeld, A.J.; Kim, S.C. Large Database Research Discussion Group Optimum Designs for Large Database Research in Musculoskeletal Pain Management. *J. Bone Joint Surg. Am.* **2020**, *102* (Suppl. S1), 54–58. [[CrossRef](#)] [[PubMed](#)]
21. Fernandez, J.; Dickinson, A.; Hunter, P. Population Based Approaches to Computational Musculoskeletal Modelling. *Biomech. Model. Mechanobiol.* **2020**, *19*, 1165–1168. [[CrossRef](#)] [[PubMed](#)]
22. Jena, P.K. *Impact of COVID-19 on Higher Education in India*; Social Science Research Network: Rochester, NY, USA, 2020.
23. MacDonald, C.W.; Lonnemann, E.; Petersen, S.M.; Rivett, D.A.; Osmotherly, P.G.; Brismée, J.M. COVID 19 and Manual Therapy: International Lessons and Perspectives on Current and Future Clinical Practice and Education. *J. Man. Manip. Ther.* **2020**, *28*, 134–145. [[CrossRef](#)] [[PubMed](#)]
24. Rahmad, N.A.; As'ari, M.A.; Ghazali, N.F.; Shahar, N.; Suffri, N.A.J. A Survey of Video Based Action Recognition in Sports. *IJEECS* **2018**, *11*, 987. [[CrossRef](#)]
25. Upper Limb Tension Tests as Tools in the Diagnosis of Nerve and Plexus Lesions: Anatomical and Biomechanical Aspects—ScienceDirect. Available online: <https://www.sciencedirect.com/science/article/abs/pii/S026800339900042X> (accessed on 12 September 2021).
26. Clinical Neurodynamics—1st Edition. Available online: <https://www.elsevier.com/books/clinical-neurodynamics/shacklock/978-0-7506-5456-2> (accessed on 12 September 2021).
27. Kleinrensink, G.J.; Stoeckart, R.; Mulder, P.G.; Hoek, G.; Broek, T.; Vleeming, A.; Snijders, C.J. Upper Limb Tension Tests as Tools in the Diagnosis of Nerve and Plexus Lesions. Anatomical and Biomechanical Aspects. *Clin. Biomech.* **2000**, *15*, 9–14. [[CrossRef](#)]
28. *Musculoskeletal Examination*, 4th ed.; Wiley: Hoboken, NJ, USA, 2015; Available online: <https://www.wiley.com/en-ad/Musculoskeletal+Examination%2C+4th+Edition-p-9781118962763> (accessed on 12 September 2021).
29. Reliability of Elbow Extension, Sensory Response, and Structural Differentiation of Upper Limb Tension Test A in a Healthy, Asymptomatic Population—IOS Press. Available online: <https://content.iospress.com/articles/physiotherapy-practice-and-research/ppr190130> (accessed on 12 September 2021).

30. The Effectiveness of Tensioning Neural Mobilization of Brachial Plexus in Patients with Chronic Cervical Radiculopathy: A Randomized Clinical Trial—ProQuest. Available online: <https://www.proquest.com/openview/fd14d5270653689800cd7e2841441855/1?pq-origsite=gscholar&cbl=2026484> (accessed on 12 September 2021).
31. Concordance of Upper Limb Neurodynamic Tests With Medical Examination and Magnetic Resonance Imaging in Patients With Cervical Radiculopathy: A Diagnostic Cohort Study—ScienceDirect. Available online: <https://www.sciencedirect.com/science/article/abs/pii/S0161475413002352> (accessed on 12 September 2021).
32. Rafiq, M.; Rafiq, G.; Agyeman, R.; Choi, G.S.; Jin, S.-I. Scene Classification for Sports Video Summarization Using Transfer Learning. *Sensors* **2020**, *20*, 1702. [[CrossRef](#)] [[PubMed](#)]
33. Yasir Farhad, M.; Hossain, S.; Rezaul Karim Tanvir, M.D.; Ameen Chowdhury, S. Sports-Net18: Various Sports Classification Using Transfer Learning. In Proceedings of the 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 19–20 December 2020; pp. 1–4.
34. Enhancing Anomaly Detection in Surveillance Videos with Transfer Learning from Action Recognition | Proceedings of the 28th ACM International Conference on Multimedia. Available online: <https://dl.acm.org/doi/10.1145/3394171.3416298> (accessed on 14 November 2021).
35. Chakraborty, S.; Mondal, R.; Singh, P.K.; Sarkar, R.; Bhattacharjee, D. Transfer Learning with Fine Tuning for Human Action Recognition from Still Images. *Multimed. Tools Appl.* **2021**, *80*, 20547–20578. [[CrossRef](#)]
36. Bilal, M.; Maqsood, M.; Yasmin, S.; Hasan, N.U.; Rho, S. A Transfer Learning-Based Efficient Spatiotemporal Human Action Recognition Framework for Long and Overlapping Action Classes. *J. Supercomput.* **2021**, 1–36. [[CrossRef](#)]
37. A Web Application Supported Learning Environment for Enhancing Classroom Teaching and Learning Experiences—ScienceDirect. Available online: <https://www.sciencedirect.com/science/article/pii/S1877042812049786> (accessed on 14 April 2021).
38. Bhattacharya, S.; Shaw, V.; Singh, P.K.; Sarkar, R.; Bhattacharjee, D. SV-NET: A Deep Learning Approach to Video Based Human Activity Recognition. In *11th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2019)*; Abraham, A., Jabbar, M.A., Tiwari, S., Jesus, I.M.S., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2021; Volume 1182, pp. 10–20, ISBN 978-3-030-49344-8.
39. Wada, H.; Gabazza, E.C.; Asakura, H.; Koike, K.; Okamoto, K.; Maruyama, I.; Shiku, H.; Nobori, T. Comparison of Diagnostic Criteria for Disseminated Intravascular Coagulation (DIC): Diagnostic Criteria of the International Society of Thrombosis and Hemostasis and of the Japanese Ministry of Health and Welfare for Overt DIC. *Am. J. Hematol.* **2003**, *74*, 17–22. [[CrossRef](#)] [[PubMed](#)]
40. Sirsat, M.S.; Fermé, E.; Câmara, J. Machine Learning for Brain Stroke: A Review. *J. Stroke Cerebrovasc. Dis.* **2020**, *29*, 105162. [[CrossRef](#)] [[PubMed](#)]
41. Heo, J.; Yoon, J.G.; Park, H.; Kim, Y.D.; Nam, H.S.; Heo, J.H. Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke* **2019**, *50*, 1263–1265. [[CrossRef](#)] [[PubMed](#)]
42. Ouahabi, A.; Taleb-Ahmed, A. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. *Pattern Recognit. Lett.* **2021**, *144*, 27–34.