Journal of Biomedical Informatics 141 (2023) 104357

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

# Deep learning prediction of motor performance in stroke individuals using neuroimaging data

Rukiye Karakis [a], Kali Gurkahraman [b], Georgios D. Mitsis [c], Marie-Hélène Boudrias [d,e,*]

[a] Department of Software Engineering, Faculty of Technology, Sivas Cumhuriyet University, Turkey
[b] Department of Computer Engineering, Faculty of Engineering, Sivas Cumhuriyet University, Turkey
[c] Department of Bioengineering, Faculty of Engineering, McGill University, Montreal, QC, Canada
[d] School of Physical and Occupational Therapy, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
[e] BRAIN Laboratory, Jewish Rehabilitation Hospital, Site of Centre for Interdisciplinary Research of Greater Montreal (CRIR) and CISSS-Laval, QC, Canada

ARTICLE INFO

*Keywords:*
Deep learning
Diffusion tensor imaging
Upper-limb motor impairment
Machine learning

ABSTRACT

The degree of motor impairment and profile of recovery after stroke are difficult to predict for each individual. Measures obtained from clinical assessments, as well as neurophysiological and neuroimaging techniques have been used as potential biomarkers of motor recovery, with limited accuracy up to date. To address this, the present study aimed to develop a deep learning model based on structural brain images obtained from stroke participants and healthy volunteers. The following inputs were used in a multi-channel 3D convolutional neural network (CNN) model: fractional anisotropy, mean diffusivity, radial diffusivity, and axial diffusivity maps obtained from Diffusion Tensor Imaging (DTI) images, white and gray matter intensity values obtained from Magnetic Resonance Imaging, as well as demographic data (e.g., age, gender). Upper limb motor function was classified into "Poor" and "Good" categories. To assess the performance of the DL model, we compared it to more standard machine learning (ML) classifiers including k-nearest neighbor, support vector machines (SVM), Decision Trees, Random Forests, Ada Boosting, and Naïve Bayes, whereby the inputs of these classifiers were the features taken from the fully connected layer of the CNN model. The highest accuracy and area under the curve values were 0.92 and 0.92 for the 3D-CNN and 0.91 and 0.91 for the SVM, respectively. The multi-channel 3D-CNN with residual blocks and SVM supported by DL was more accurate than traditional ML methods to classify upper limb motor impairment in the stroke population. These results suggest that combining volumetric DTI maps and measures of white and gray matter integrity can improve the prediction of the degree of motor impairment after stroke. Identifying the potential of recovery early on after a stroke could promote the allocation of resources to optimize the functional independence of these individuals and their quality of life.

## 1. Introduction

Stroke is a neurological disorder that causes wide ranging deficits in the cognitive and motor function of survivors [1]. It is the second most common cause of death among adults and the third most common cause of disability worldwide [2]. Stroke can lead to long-term impairments such as hemiparesis or speech disabilities and affect cognitive functions, including memory [2–4]. The degree of motor impairment and potential for recovery after stroke observed for each patient are difficult to predict. Socio-demographic measures (e.g., age, gender), and clinical measures such as the National Institutes of Health Stroke Scale (NIHSS) or Fugl-Meyer Assessment (FMA) are commonly used as predictors of motor recovery early after stroke [3,5–11].

More recently, neuroimaging techniques such as structural magnetic resonance imaging (MRI) and computed tomography (CT) have also been used to identify biomarkers of post-stroke motor recovery [5,9–10,12–13] with limited accuracy up to date. For instance, diffusion metrics obtained from MRI scans provide key information about the microstructural properties of the brain's white matter (WM) by quantifying the random movement of water molecules. There are two types of diffusion imaging methods: diffusion-weighted imaging (DWI) and diffusion tensor imaging (DTI). In DWI, each voxel's diffusion rate at the microscopic level is estimated using strong magnetic field gradients. After a stroke, it has been shown that DWI is more sensitive in identifying the changes in WM content compared to structural MR images [3]. However, DWI is limited in revealing the complex diffusion patterns

* Corresponding author.
*E-mail address:* mh.boudrias@mcgill.ca (M.-H. Boudrias).

https://doi.org/10.1016/j.jbi.2023.104357
Received 15 November 2022; Received in revised form 24 February 2023; Accepted 1 April 2023
Available online 7 April 2023

occurring in the brain [3,14]. DTI, on the other hand, quantitatively estimates the direction in which the diffusion of the tissue is more restricted, providing more information about the changes in brain structure and the extent of a lesion due to stroke. In post-stroke DTI images, it has been found that diffusion is significantly reduced in WM compared to gray matter (GM) [15]. The diffusion of each WM voxel can also be used to detect anisotropy changes in stroke lesions and monitor the brain's response to treatments and motor functional recovery [3,14–16]. Using DTI tractography algorithms, motor deficits present after stroke were shown to be associated with the integrity of the corticospinal tract (CST) [17]. The quantification of a lesion's volume using MRI-based DTI was found to be a moderate predictor of the degree of the motor deficit, given that small lesions can affect critical motor regions for movement production [7].

Additional DTI metrics such as fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD), and axial diffusivity (AD) have also been used to predict motor recovery after stroke [3,5]. For instance, FA can provide an estimate of the number of fibers affected by a lesion. In particular, FA measurements of the CST can be used to evaluate motor recovery [3]. Acute ischemic stroke reduces MD in the lesion significantly, which returns to normal within a few days. AD has been more strongly associated with axonal degeneration, while RD has been related to myelination. This suggests that combining diffusion measurements may help develop a more effective method for estimating motor recovery in patients at different stroke stages, as well as with different lesion sizes and locations.

Grey matter (GM) integrity can also be used as a predictor of motor impairment after stroke [18–20]. Using a general linear model, a voxel-level false discovery rate, and region of interest (ROI)-based analyses, Diao *et al.* [18] investigated GM volumes (GMV) in individuals in the chronic phase of a subcortical stroke. A GMV increase was observed in bilateral paracentral lobules, the supplementary motor area (SMA), and the right middle occipital gyrus for those with a right lesion when compared to healthy individuals. A correlation between motor recovery and GMV increase was found in the bilateral SMAs in these patients. In another study by Dang *et al.* [19], changes in GMV in the ipsi- and contralesional SMA correlated with clinical scores (FMA & Barthel index) for individuals with focal cerebral infarct. In the study of Yang *et al.* [20], atrophy in regions of the brain far from the infarct in the same hemisphere was associated with motor impairment present after stroke. The volume of interest ratios of motor-related brain regions and the temporal lobe between the ipsilesional and contralesional hemispheres, as well as the FA ratio of the entire CST, were found to be correlated with motor impairment. These studies suggest that adding GM volumetry to DTI-based measures could help better predict the extent of motor deficits observed in stroke individuals.

Some studies have attempted to combine clinical, neurophysiological, and neuroimaging measures to predict motor outcomes after stroke. For instance, Stinear *et al.* [12] developed the PREP algorithm, which combines clinical motor assessments as well as transcranial magnetic stimulation (TMS) and FA-based measures of CST integrity, to predict the recovery of upper limb function after stroke. The accuracy, specificity, and sensitivity values obtained from their algorithm reached 0.64, 0.88, and 0.73, respectively. In another study from the same group, the PREP2 algorithm was subsequently introduced to categorize the upper limb motor outcomes three months after stroke into four groups: "Excellent", "Good", "Limited" and "Poor". Adding two T1-weighted MRI measures, i.e., the size of the lesion, and measures of the integrity of the CST and sensorimotor tracts to their algorithm slightly improved the positive predictive value (PPV) and overall accuracy to 0.78 and 0.75 with shoulder abduction and finger extension (SAFE) score higher than 5 when using a classification and regression tree (CART) analysis [13]. Thus, combining measures from different sources can help the prediction of the recovery profile of stroke patients. However, the PREP2 algorithm reaches a limited accuracy of 0.70, particularly for the prediction of more impaired patients with SAFE score less than 5 [13].

Machine learning (ML) allows the combination of different features to classify the labelled samples. It has been successfully used to classify neurological diseases such as Alzheimer's disease, epilepsy, schizophrenia, stroke, autism, and mild cognitive impairment [21–22]. ML classifiers can be used to estimate motor recovery after stroke [23–25]. However, they require feature analysis, and the most informative features to this end are still unknown. Deep Learning (DL) is an ML method based on artificial neural networks with a large number of layers that is able to learn complex patterns that are present in the input data [26–27]. It is not straightforward to explain how the DL selects the features that are useful for classification, due to the deep architecture of the model. However, DL models reduce human effort in feature engineering compared to traditional ML methods using handcrafted features. Specifically, Convolutional Neural Network (CNN) is the DL architecture that has achieved the highest accuracy value in image classification [26–27]. CNN has the advance of not requiring feature extraction, unlike traditional ML techniques used in the context of medical radiology, such as support vector machines (SVM), Random Forests (RF), k-nearest neighbor classification (k-NN), and Naïve Bayes (NB) [28–32]. DL has been utilized in the context of medical image segmentation [33–34] using CT, MRI, or DWI images, reaching performance values ranging between 0.30 and 0.87 according to the Dice coefficient. DL has also been used in the context of stroke classification [35–36]. Nielsen *et al.* [35] used MR images of acute ischemic stroke participants to predict the treatment effect of a recombinant tissue-type plasminogen activator using a CNN model with an area under curve (AUC) of 0.88. Heo *et al.* [36] used 3 ML approaches including deep neural networks (DNN), RF, and logistic regression (LR) to predict the long-term motor outcomes of acute ischemic stroke individuals using the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score. The AUC values of the DNN, RF, and LR models were 0.88, 0.86, and 0.85, respectively.

DTI, measures of WM, and GM integrities obtained from MRI images can thus be useful to determine the extent of a brain lesion. On the other hand, it is difficult to establish the precise motor recovery profile using these metrics post-stroke. In addition, determining which of the aforementioned markers is most relevant is challenging, as using multiple metrics increases the correlation's reliability. In this context, the present study aimed to predict motor impairment by using DL techniques, which have great potential for classifying patients without using hand-crafted features, using DTI and structural MR images of stroke patients. Here a large dataset, including 154 structural images of 123 individuals was used. The hypothesis was that a combination of demographic data and brain imaging measures such as FA, AD, MD, RD, GM, and WM incorporated within a multi-channel 3D-CNN using residual blocks would improve the prediction of motor impairment observed post-stroke.

## 2. Statement of significance

| Problem or Issue | What is Already Known | What this Paper Adds |
|---|---|---|
| The degree of motor impairment and the profile of recovery after stroke are difficult to predict for each individual post-stroke. | Measures obtained from clinical assessments, neurophysiological, and neuroimaging techniques have been used as potential biomarkers of motor recovery, with so far limited accuracy. Determining which of these biomarkers is most relevant is challenging as using multiple metrics increases the correlation's reliability. | The present study proposes a multi-channel 3D DL model with residual blocks combining volumetric DTI maps and measures of white and gray matter integrity to predict motor impairment. It achieved higher accuracy than traditional ML techniques in which feature selection can be challenging and time consuming. In addition, using whole-brain MRI and DTI metrics in DL |

(*continued on next page*)

(*continued*)

| Problem or Issue | What is Already Known | What this Paper Adds |
|---|---|---|
| | | has been shown to extract richer features than simply using ROI analysis. |

## 3. Materials and methods

In the proposed DL system, DTI and MR images, as well as demographic data, were used as inputs for the multi-channel 3D-CNN model. Motor impairment was classified into 2 categories, "Poor" and "Good". The classification of motor impairment was performed with DL and other ML classifiers. ML classifiers such as k-NN, SVM, Decision Trees (DT), RF, Ada Boosting (AB), and NB, the inputs of which were obtained from the fully connected layer (FCL) of the CNN, were used to compare performance with the DL model. The outputs of the trained models and the outputs labeled by experts were compared based on accuracy, specificity, precision, recall, and F1 values obtained from Receiver Operating Characteristics (ROC) analysis.

### 3.1. Dataset

Three different datasets were used to test the ability of the multi-channel 3D-CNN model to predict motor impairment in healthy and stroke individuals. The details of the demographic information and motor scores of participants are provided in Table 1. The first dataset included a total of 42 participants (19 S patients, and 23 healthy controls) [4]. Motor impairment was calculated as a percentage of the score obtained from the unaffected hand based on three tests including hand grip strength (GS) [37], Nine Hole Peg Test (NHPT) for finger dexterity [38], and Box and Block Test (BBT) for unilateral gross manual dexterity [39]. DTI was collected using a 3 T Siemens Trio TIM scanner at the Wellcome Centre for Human Neuroimaging at University College London with a single-shot diffusion-weighted echo-planar imaging (*EPI*) sequence (61 non-collinear directions, b = 1000 s/mm$^2$, TE/TR = 102 ms/182 ms, FOV = 128 × 128, 85 slices, and voxel size = 1.72 × 1.72 × 1.7 mm$^3$) plus 7 volumes with low diffusion weighting (b = 100 s/mm$^2$) [4]. A T1-weighted structural scan (TE/TR = 2.48 ms/7.92 ms, flip angle = 16°, FOV = 240 × 256 mm, 176 slices, and voxel size = 1 × 1 × 1 mm$^3$) was also collected.

The second dataset included a total of 50 participants (14 S patients, and 36 healthy controls) [40]. The motor impairment of participants was determined using the Fugl-Meyer Assessment (FMA) upper-limb score [41]. T1-weighted MR (TE/TR = 3.65 ms/7.47 ms, flip angle = 6°, FOV = 256 × 256 mm, 165 slices, and voxel size = 1 × 1 × 1 mm$^3$) and DTI data were acquired at the University of British Columbia MRI Research Centre using a 3 T Philips Achieva scanner. DTI images were collected using a single-shot *EPI* sequence (60 non-collinear directions, b = 700 s/mm$^2$, TE/TR = 60 ms/7096 ms, FOV = 112 × 112 mm, 70 slices, and voxel size = 2.0 × 2.0 × 2.2 mm$^3$) with a single unweighted volume (b = 0 s/mm$^2$) [40].

The third dataset includes a total of 31 participants, for whom DTI and MRI scans were acquired using a 3.0 T Siemens Trio TIM scanner at the University of Geneva [42]. The motor impairment of participants was measured using the FMA upper-limb score [42]. *EPI* diffusion-weighted volumes (30 non-collinear directions, b = 1000 s/mm$^2$, TE/TR = 82 ms/8.2 s, FOV = 128x128, 64 slices, voxel size = 1.8x1.8x2.0 mm$^3$) were acquired with one unweighted volume (b = 0 s/mm$^2$). T2-weighted MR scans (TE/TR = 376 ms/5 s, FOV = 512 × 512 mm, 176 slices, voxel size = 0.45 × 0.4 5x 0.90 mm$^3$) volumes were also acquired [42].

In the present study, the FMA scores of participants were used as the ground truth of the DL model. However, for the first dataset shown in Table 1, three different behavioral measurements (BBT, GS, and NHPT) were obtained for each participant. For this reason, a single representative motor score was calculated using principal component analysis (PCA). PCA has the advantage of preserving as much variability (i.e., statistical information) as possible by projecting each data point into the first few principal components to reduce the dimensionality of a dataset. The first principal component yields the direction of maximum variance for the data. This technique has been used by Rondina et al. [23,43] to convert four assessment scores, including the Action Research Arm Test (ARAT), Motricity Index, GS, and NHPT, into a single representative score to predict motor outcomes [43] and classifying motor recovery as "Good" and "Poor" [23]. PCA has also been used to analyze sub-items of a motor score [44]. In the present study, the first principal component was calculated using three scores (BBT, GS, and NHPT). The mean of each test score was calculated and subtracted from the test score vector (X) resulting in the matrix Y. Then, the test score values were normalized by dividing each element of Y by the square root of the covariance of X.

**Table 1**
Demographic information and motor scores of participants.

| Name | Number of Individuals | Stroke Time (Month) | Age (SD) | Gender (M/F[a]) | Lesion Side (R/L[b]) | Hand Affected (R/L[b]) | BBT[c] (%) | NHPT[d] (%) | GS[e] (%) | FM[f] (%) | G/P[g] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset-I (42) | Healthy (23) | – | 47 (17.9) | 8 /15 | – | – | 100.7 (9.8) | 96.9 (10.5) | 95.1 (8.2) | – | 23/- |
| | Stroke (19) | 39.7 (53.0) [1 209] | 52 (14.5) | 4 / 15 | 8 / 11 | 11 / 8 | 58.3 (33.6) | 43.8 (36.4) | 57.2 (29.9) | – | 11/8 |
| Dataset-II (50) | Healthy (36) | – | 65 (7.9) | 11/25 | – | – | – | – | – | 66.0 (0.0) | 36/- |
| | Stroke (14) | 79.7 (47.5) Y | 68 (11.3) | 11/3 | 9/4 | 4/9 | – | – | – | 54.1 (14.5) | 12/2 |
| Dataset-III (62) | Stroke-Pre* (31) | 0.5 | 64 (12.9) | 17/14 | 20/11 | 11/20 | – | – | – | 11.9 (9.9) | 4/27 |
| | Stroke-Post** (31) | 3 | 64 (12.9) | 16/15 | 21/10 | 10/21 | – | – | – | 29.7 (22.2) | 15/16 |

[a] M: Male-F: Female.
[b] R: Right-L: Left.
[c] BBT: Box and Block Test.
[d] NHPT: Nine-Hole Peg Test.
[e] GS: Grip Strength.
[f] FM: Fugl Meyer Test.
[g] G/P: Good/Poor.
* Pre: clinical assessment at 2–4 weeks.
** Post: clinical assessment at three months.

Finally, the eigenvalue vector of the normalized values was obtained by singular value decomposition. Thus, the first column of the resulting matrix provided the values of the patients' first principal component. These principal components were used to obtain the cutoff point of two recovery profiles using descriptive statistics such as average and median [4,23,43].

In the present study, the FMA scores of datasets II and III, as well as the PCA scores of dataset I were classified into 2 categories of recovery profiles, "Poor" and "Good" by the DL. In the literature, there are different FMA cutoff values to categorize the severity of patients [44,45]. Here, the cutoff value was established by evaluating the histogram, mean, and median analysis of the FMA and PCA scores. The mean FMA and median values for the datasets II-III were 26.7 and 20.0, respectively. This agrees with the categorization used by Woytowicz et al. [44], where a severe class cutoff value of 27 was determined for the individuals with low motor recovery potential. This value allowed to divide the patients with upper-limb motor deficits into two classes, as shown in the histogram distribution (Fig. 1b). FMA scores were unavailable for all three different datasets in this study. Since Dataset-I included three other scores (BBT, GS, and NHPT), these measures were converted into a single representative score using PCA, as done in Rondina et al. [23,43]. To our knowledge, there is no method in the literature to map these measures (BBT, GS, and NHPT) onto the FMA score. The cutoff value for PCA results was also selected using the mean value, which was found to be very close to 0. Therefore, the patient with motor impairment in Dataset-I was assigned to the Poor class when its first principal component was $\leq 0$. The histogram distributions and the cutoff values for the datasets are given in Fig. 1a. Healthy controls had normal motor performance, and were assigned to the "Good" class.

### 3.2. Deep learning method

A multi-channel 3D-CNN model was used to evaluate the degree of upper-limb motor impairment after stroke. The analyses consisted of three phases. Initially, FSL (*https://fsl.fmrib.ox.ac.uk/fsl/*) and custom Matlab scripts (MATLAB 2016b, The MathWorks, Inc., Natick, Massachusetts, United States) were used to preprocess DTI and T1/T2 weighted MR images. The preprocessing included noise removal, obtaining DTI maps, as well as obtaining GM and WM images. As a second step, DTI-derived maps such as FA, AD, RD, and MD, T1/T2 weighted MR images maps of GM and WM, and demographic data (age, gender, etc.) were used as input in the multi-channel 3D-CNN architecture, as shown in Fig. 2 [46–47]. Finally, once the training of the DL model was completed, the k-NN, SVM, DT, RF, Ada Boosting (AB), and NB classifiers were trained and tested to determine the best classifier. The input features of ML classifiers were obtained from the flatten, first, or second FCL of the CNN model. The results were compared based on the AUC, accuracy, specificity, precision, recall, and F1 values obtained from the Receiver Operating Characteristics (ROC) analysis.

*Preprocessing of DTI data:* For each subject, noise was removed from the DTI images (dwidenoise by MRtrix, *https://www.mrtrix.org/*). Distortion and motion correction was performed to remove eddy currents that are a natural effect of the changing magnetic field on diffusion scans (eddy_correct implemented in FSL). After the motion correction process, which caused a significant bias in diffusion measurements and orientation, the b-vectors were rotated to preserve the correct orientation information [47–48]. The brain was extracted from the skull and non-brain areas using the BET tool in FSL. Diffusion tensor and diffusivity metrics were calculated using the dtifit function implemented in FSL and MATLAB to compute diffusion maps of FA, AD, RD, and MD. Each map was registered to a DTI template using a non-linear registration approach (FLIRT + FNIRT in FSL). DTI maps of the subjects with lesions on the left hemisphere were flipped from left to right using the fslswapdim function in FSL to ensure that subjects had their lesion on the same side for direct comparison.

*Preprocessing of T1/T2 weighted MRI data:* Using FSL, the MR images were oriented for each subject using the command fslreorient2std. MR images were denoised using BM4D filters with MATLAB, the details of which can be found in [49–50]. The skull was stripped using BET in FSL. Bias correction was performed by the FAST FSL method. MR images were co-registered to the MNI 2 mm template by using non-linear registration (FLIRT + FNIRT in FSL). For patients with left hemisphere lesions, the images were flipped to ensure all the patients' lesions were on the right side. Finally, MR images were segmented into WM, GM, and cerebrospinal fluid (CSF) using the FAST method in FSL.

*Masks of Motor Regions:* Regions of interest (ROI) were chosen based on their known roles in motor and sensory functions [23,43,51–52], as shown in Fig. 3. Motor areas included the pre- and postcentral gyrus, SMA, superior frontal gyrus, middle frontal gyrus, inferior and superior parietal regions, thalamus, caudate, putamen, and pallidum. The CST was also selected as the extent of its damage after stroke can provide insight into upper-limb function [23,43] and potentially increase the accuracy of ML classification. Masks were created using the Automated Anatomical Labeling (AAL) atlas30 [53] implemented in MATLAB, with an isotropic voxel size of 2x2x2 mm. The CST mask was derived from the Johns Hopkins University (JHU) WM tractography atlas implemented in the FSL toolbox. Finally, the remaining brain structures were removed by multiplying each ROI mask with the images. Two DL models were trained using the whole brain and ROI images, including motor regions and the CST.

The DTI maps and structural images of each participant were entered as 6 separate channels with the same dimension ($91 \times 109 \times 91$) in the multi-channel 3D-CNN model. The dataset was divided into 80 % training and 20 % testing sets to compare the accuracy of the DL models.

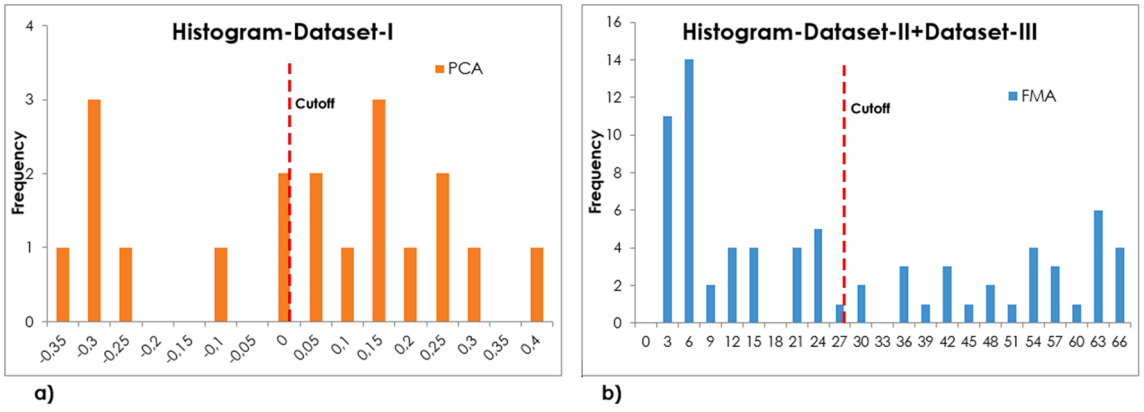


**Fig. 1.** Histogram distributions in the employed datasets for a) PCA values obtained from the three motor scores (BBT, GS & NHPT; Dataset I), b) FMA scores (Datasets II and III).

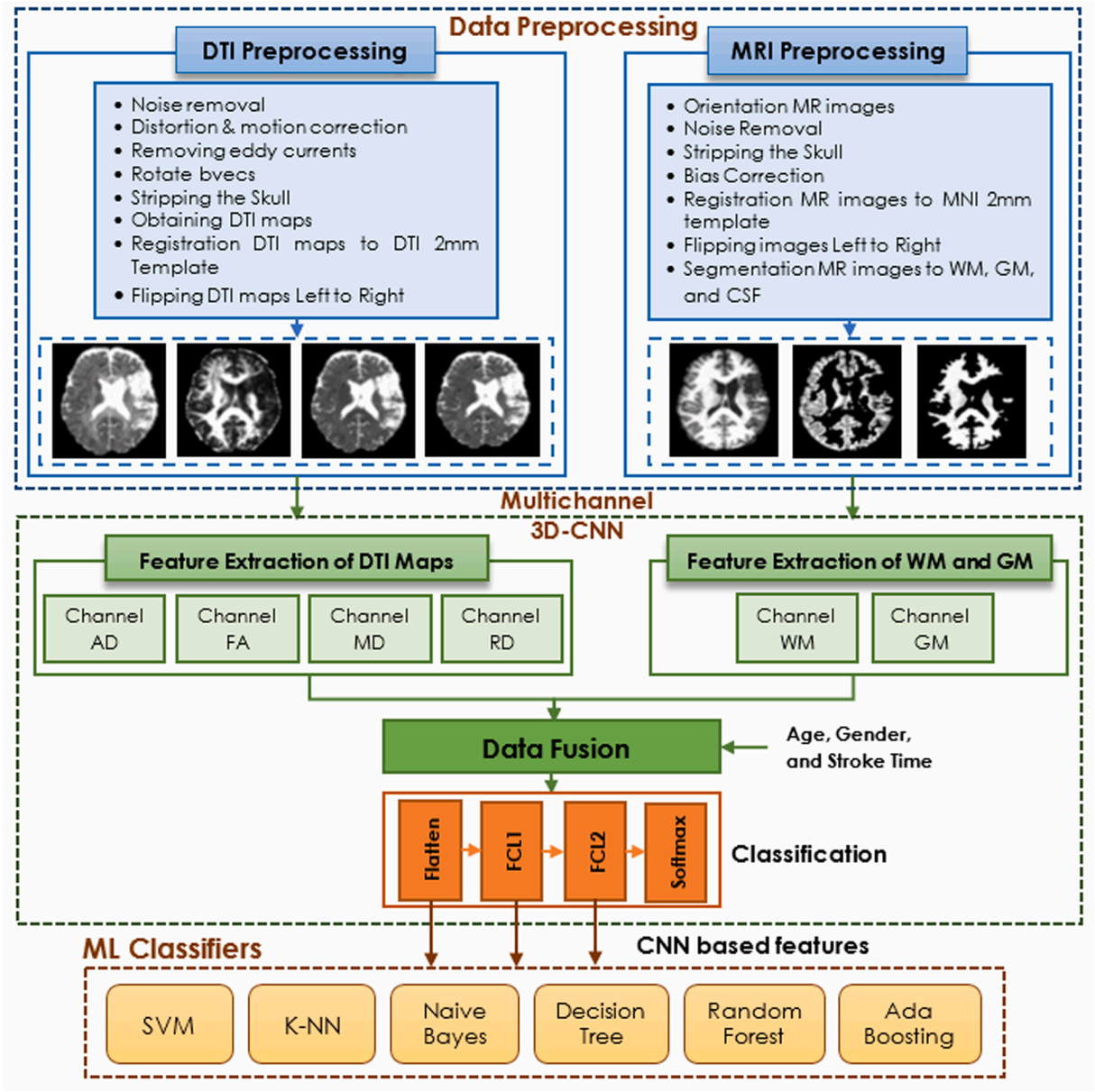


**Fig. 2.** Flowchart for the comparison of the proposed multi-channel 3D-CNN-based models.

The required libraries and toolboxes needed for the DL analysis, such as Keras, Tensorflow, scikit, pandas, OpenCV, nibabel, and dipy were obtained via open sources. The proposed multi-channel 3D-CNN model was coded using Python 3.7. The MATLAB platform was used for the performance analysis.

### 3.2.1. 3D convolution neural network (CNN)

The CNN consists of multiple building blocks such as convolution layers, pooling, and FCLs. It automatically and adaptively learns the properties of the input data through backpropagation. The first two layers of the CNN are the convolution and pooling layers, which form the feature maps. Convolution is a specific type of linear operation, which allows the network to detect important features at given spatial positions in the input using filters. The convolution process preserves the relationship between voxels by learning the features of the image [22,26–29].

The 3D-CNN architecture uses 3D convolution to obtain features along both the spatial and temporal dimensions by dividing the volume into small cubes. The 3D convolution denoted as $h$ in the *i-th* feature volume map of the *l-th* layer was calculated using Eq. (1) [54].

$$u_{ki}^{l}(x,y,z) = \sigma\left(\left(\sum_{k} h_{k}^{l-1}(x-m, y-n, z-t) * W_{ki}^{l}(m,n,t) + b_{i}^{l}\right)\right) \quad (1)$$

where $h_k^{l-1}$ is the *k-th* 3D feature volume map of the previous layer. $W_{ki}^{l}(m,n,t)$ and $b_i^l$ are the weight in the 3D convolution kernel at position *(m, n, t)* and bias term, respectively. $\sigma(.)$ is the non-linear activation function [54].

The 3D feature map obtained by convolution is transformed into a nonlinear map through an activation function such as sigmoid, tanh, or ReLU. A subsampling process is performed in the pooling layer, reducing the size of the feature maps and the number of the network's learnable parameters. The average or maximum pooling function is commonly preferred, and there are no learnable parameters in the pooling layer [22,26–28].

The FCLs that flatten the feature maps and transform them into a single-dimensional number array are used after the last convolution or pooling layer [22,26–28]. The output of a neuron in the FCL is connected with a different weight to all inputs of the next FCL. In each neuron of

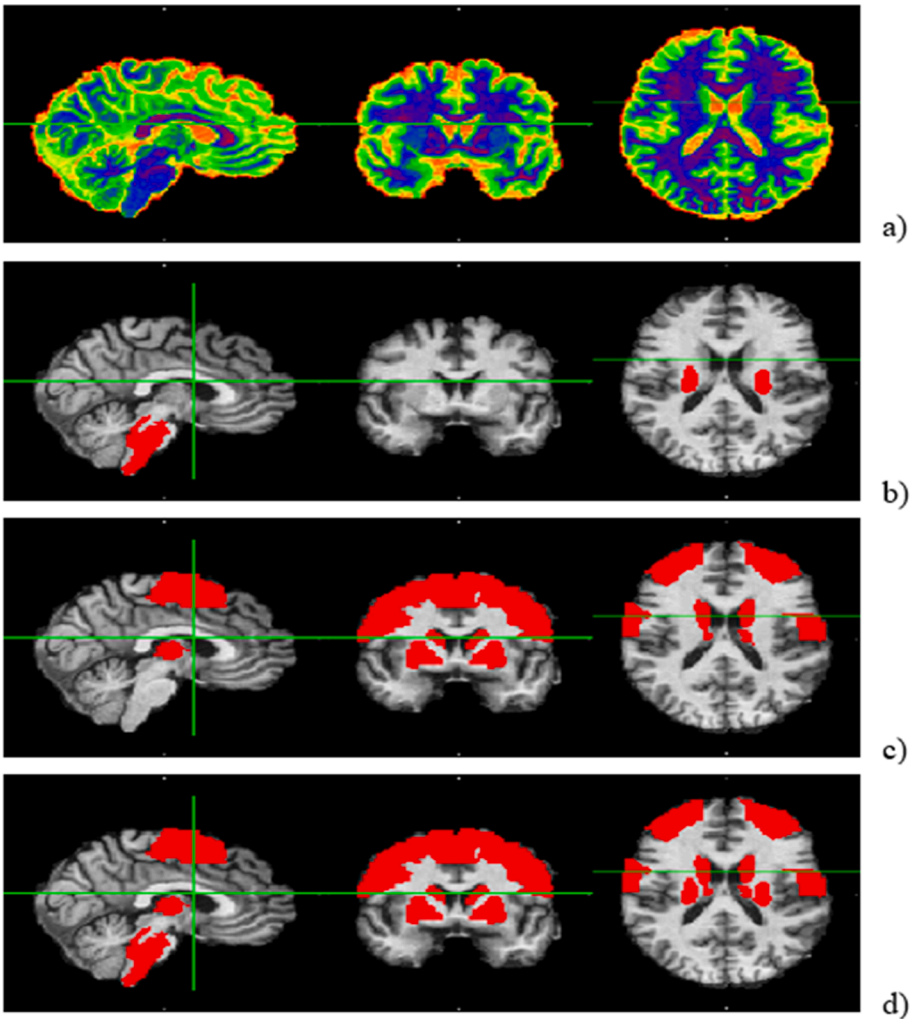


**Fig. 3.** Masks used as inputs in the DL architecture (a) whole-brain, (b) CST, (c) motor regions, (d) CST and motor regions.

the FCL, the input information is first calculated, and then an activation function such as ReLU is used. The last layer has the same number of neurons as the ground-truth label number of the problem, and the probabilities of the outputs are usually calculated with the softmax function.

The CNN was trained by minimizing the loss function given in Eq. (2).

$$\mathscr{L} = \frac{1}{N}\sum_{n=1}^{N} l(\theta; y^{(n)}, o^{(n)}) \tag{2}$$

where $y^{(n)}$ and $o^{(n)}$ are the target output and the predicted output of the network for *n-th* input data $x^{(n)}$, respectively. $\theta$ represents all the learnable parameters of the network such as weights and bias values. $l(.)$ is the loss calculated from the target and predicted outputs for each sample.

### 3.2.2. Deep residual learning model

The main factor that causes variations in the classification DL architecture is the feature extraction layers. VGGNet, ResNet, DenseNet, and InceptionResNet-based architectures have been widely used as the feature extraction layers of DL models in 3D multimodal medical image analysis [55–56]. VGGNet differs from traditional CNN models in terms of the size and number of filters in sequential convolution blocks. Therefore, increasing the number of blocks dramatically increases the number of parameters. The high number of trainable parameters requires large data amounts, because it causes a forgetting problem in the network. As an alternative to traditional architectures, the ResNet, DenseNet, and InceptionResNet architectures, which are deeper but with smaller filter sizes in the convolutional layers can be used. These architectures use residual links that transfer gradients between blocks to overcome the forgetting problem. Since DenseNet and InceptionResNet models have a deeper architecture, computational time and memory consumption are higher than ResNet for multimodal 3D image analysis. Although there is no comparison specific to 3D medical image analysis, based on Bianco et al. [57], ResNet50 has a higher memory consumption and extraction time, but it also yields higher accuracy in the case of 2D image analysis. In addition, Suganyadevi et al. [58] examined 120 medical image analysis studies and concluded that the ResNet architecture yielded the best overall performance. In the present study, the selection of DL architecture was based on two criteria. The first was the characteristics of the data to be processed, and the second was hardware-related limitations. The data used are 3D, multimodal and have different sizes of lesions. For this reason, it was assumed that the classification performance would be better, as the 3D-DL architectures would preserve the volumetric information in the processing of 3D neuroimaging data. The relationship between the location of the lesion with motor recovery can be more important than the size of the lesion [7]. Therefore, classifying motor impairment from stroke data is a complex problem. It is necessary to use architectures with a larger number of deep layers to process such data. Accuracy can also saturate to a certain level, due to the vanishing gradient problem that occurs in

deep architectures. For this reason, the ResNet50 architecture with skipping connections was used to avoid feature loss in deeper layers. Fig. 4 shows the architecture for one of the channels in the proposed 3D-CNN with residual blocks, which solves the vanishing gradient problem in deeper models [59].

In the convolutional layers, the output of layer $\ell$ is connected to the input of the layer $(\ell+1)$. This transition between layers is expressed by the $x_\ell = H_\ell(x_{\ell-1})$ formula. In the ResNet, a skip-connection is used to add an identity function $(x_{\ell-1})$ to the output of layer $\ell$, as given in Eq. (3) [59].

$$x_\ell = H_\ell(x_{\ell-1}) + x_{\ell-1} \tag{3}$$

Each channel first transferred the 3D images to the convolutional layer and then to the pooling layer. Feature vectors obtained from the last pooling layer of each channel were merged in the fully connected layer. Finally, the images transferred forward from the input layers to the classifier layer were classified into two groups (Figs. 2 & 4). In the training process, the weights and bias values of neurons between layers in the network were updated to minimize the error between real and predicted values by an optimizer method such as stochastic gradient descent (SGD), ADAM, and RMSProp [22,26–28].

### 3.2.3. Implementation

The proposed multi-channel 3D-CNN model was implemented using the Keras library in Python based on Tensorflow. The experiments were conducted on a PC with a NVIDIA Tesla K80 24 GB GPU, Intel XEON ES-2680 V4 @ 2.40 Hz CPU, and 128 GB RAM. SGD was used to train the 3D-CNN models. The momentum, learning rate, and weight decay were adjusted to 0.9, 0.0001, and 0.005, respectively. Dropout and L2 regularization were adopted to avoid overfitting. Due to memory limitations, 2 mm registered images were used in the multi-channel architecture. A single-channel is a 3D-CNN that classifies using only one MRI or DTI metric. The single-channel with 2 mm registered images of the two motor recovery profiles was trained to reveal the correlation between each map and motor scores.

There is no clear guide on the best train-test split ratio of a dataset for DL application. In the present study, 0.8–0.2, which is one of the most widely preferred experimental train-test ratio in the DL medical image analysis literature [60–61], was used. There was a total of 154 samples, 101 and 53 samples for the Good and Poor classes respectively, in our dataset. First, these samples were split into 124 training and 30 testing data using 10 repeated random subsampling validation (RSV) method. There were equal numbers of poor and good instances in the test data. Using RSV ensured the chance of selecting the same instance of Good and Poor classes at least once and twice in the subset, respectively. Then, to ensure class balance in the training dataset, data augmentation with sharpening was applied to the Poor class for each subset. In addition, the DL model was separately trained 10 times using each subset; in other words, a total of 100 training iterations were performed for the entire validation process.

### 3.2.4. Evaluation metrics

Results were compared according to the AUC, accuracy, specificity, precision, recall, and F1 values obtained from ROC analysis. A confusion matrix is a table that groups True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) to evaluate the results of ML classifiers [62]. TP means that ground truth and prediction are both positive. FP means that the ground truth is positive while the prediction is negative. TN means that the ground truth is negative, and the prediction is negative, whereas FN means that the ground truth is negative and the prediction is positive. The specificity shows that a classifier can truly identify non-positive cases (Specificity = TN / (TN + FP)). The accuracy measures the success of classifying positive cases as positive and negative samples as negative:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

The precision, given in Eq. (5), shows how many positive examples are correctly predicted.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

Recall, also termed sensitivity, measures the proportion of correctly classified positives, as shown in Eq. (6).

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

F1 represents the harmonic mean of the precision and recall values,

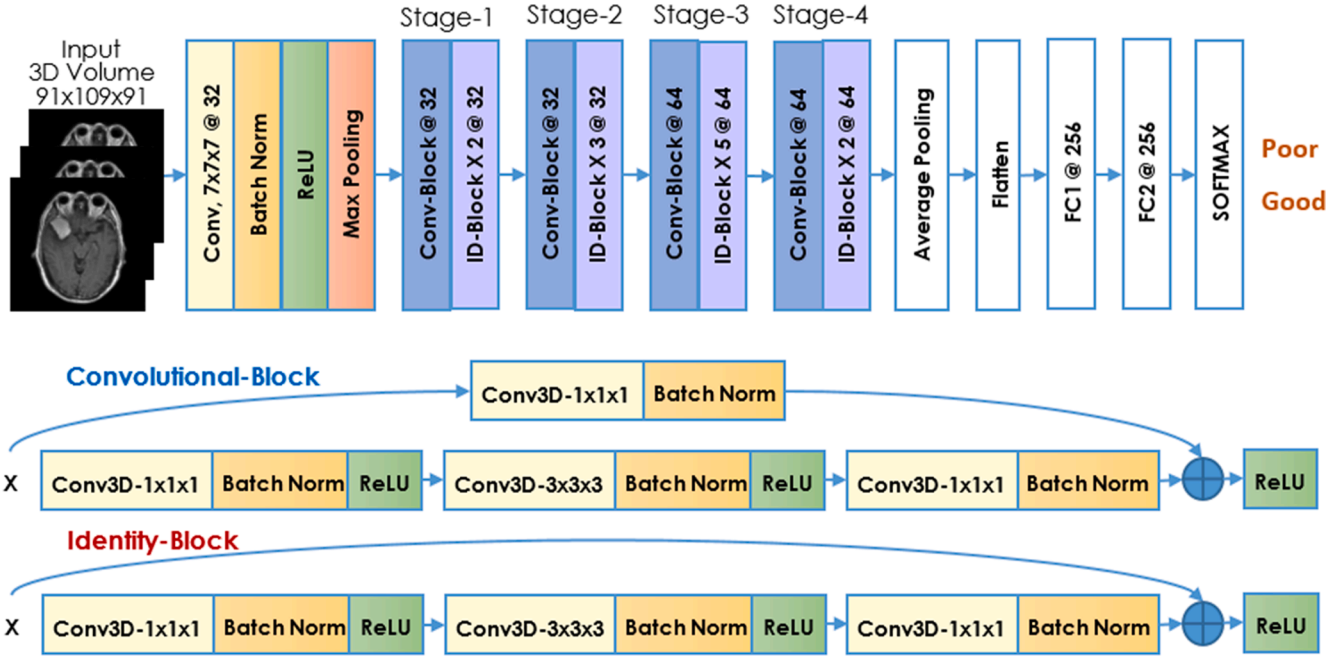


**Fig. 4.** The architecture of a single-channel of the 3D-CNN architecture with residual blocks.

and can be calculated using Eq. (7):

$$F1 = 2\frac{Precision.Recall}{Precision + Recall} = 2\frac{TP}{2TP + FP + FN} \quad (7)$$

## 4. Results

Table 2 shows the performance results of 3D-CNNs with single-channel for 10 RSV subsets. The accuracy values were between 0.837 and 0.877 for the testing data, as shown in Table 2. The DTI-based maps, namely WM, MD, and FA, yielded the best performance value, while GM images yielded the worst value. Fig. 5 illustrates the precision, recall, and F1 results of the CNN networks trained with DTI maps, WM, and GM separately. The ROC values for the DTI maps, WM, and GM were very similar (see Fig. 5 & Table 2). For this reason, the multi-channel 3D-CNN model was trained with all images (DTI maps + WM + GM) to classify upper-limb motor impairment.

The multi-channel (DTI maps, WM, and GM) 3D-CNN was trained and tested with whole-brain images by 10 RSVs × 10 training iterations, as shown in Table 3. The *overall Mean* in Table 3 corresponds to the average values of the 100 training datasets, while the *Mean of Maxs* corresponds to the mean of the maximum values obtained from the 10 subsets. The performance values of the DL model were found to be similar for both classes, as desired.

The performance results of the multi-channel 3D-CNN for whole-brain images based on the 2 motor recovery profiles are given in Table 3. All values were very similar for the 2 classes, with the Good class yielding slighter better results than the Poor class. In this study, since the Good class included healthy volunteers, it can be assumed that healthy control images were successfully classified.

Motor region maps were subsequently used as inputs of the multi-channel 3D-CNNs to classify the upper-limb motor impairment. Table 4 illustrates the test results of the models for 10 RSVs × 10 training times as in the case of whole-brain training. Considering all the performance results in the Table 4, it can be concluded that the DL model cannot distinguish the two classes equally, as seen in specificity, precision, and recall values.

In Fig. 6, the box-scatter plots of the CNN models trained and tested with the whole-brain and motor region images by 10 RSVs × 10 repetitions are given. According to the comparison of CNN models using the accuracy, specificity, and F1 values in Fig. 6, the 3D-CNN models trained with whole-brain images obtained the best results, as also observed in Tables 3 & 4.

**Table 2**
Performance results of the single-channel CNN with single-channel for 10 RSVs.

| Channel | Accuracy | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|---|
| **AD** | 0.860 [0.80 0.93]* | 0.860 [0.80 0.93] | 0.871 [0.81 0.94] | 0.860 [0.80 0.93] | 0.859 [0.80 0.93] |
| **FA** | 0.870 [0.77 0.93] | 0.870 [0.77 0.93] | 0.885 [0.78 0.94] | 0.870 [0.77 0.93] | 0.868 [0.76 0.93] |
| **MD** | 0.877 [0.83 0.93] | 0.877 [0.83 0.93] | 0.887 [0.85 0.94] | 0.877 [0.83 0.93] | 0.876 [0.83 0.93] |
| **RD** | 0.863 [0.70 0.90] | 0.863 [0.70 0.90] | 0.879 [0.76 0.90] | 0.863 [0.70 0.90] | 0.861 [0.68 0.90] |
| **WM** | 0.880 [0.80 0.97] | 0.880 [0.80 0.97] | 0.888 [0.80 0.97] | 0.880 [0.80 0.97] | 0.879 [0.80 0.97] |
| **GM** | 0.837 [0.73 0.90] | 0.837 [0.73 0.90] | 0.845 [0.75 0.90] | 0.837 [0.73 0.90] | 0.835 [0.73 0.90] |

* Mean [Minimum Maximum].

### 4.1. Results of other machine learning classifiers

The results of the DL model trained on whole-brain images were compared with those obtained using alternative classifiers including k-NN, SVM, DT, RF, AB, and NB. Since DTI maps and MR images were used as inputs, images from the last convolution and pooling layers of the DL architecture were used instead of performing complex feature analysis. In the DL architecture given in Figs. 2 & 4, images and demographic data taken from the last pooling layer of 6 different channels were flattened. These 3073 features obtained from the flatten layer, and 256 features from the first and second FCLs were separately used as inputs of the ML classifiers.

To further illustrate the satisfactory performance of the proposed method, we examined whether the employed data were statistically representative of the classes. Fig. 7a, 7b, and 7c show the *t*-SNE (Stochastic Neighbor Embedding) visualizations of the features obtained from the flatten, first, and second FCLs of the 3D-CNN model, respectively. The distribution in Fig. 7c shows that Poor features appear to be clustered more clearly. Therefore, it is concluded that the second FCL improved the distribution compared to the previous layers. Overall, the clustering patterns in the figure suggest that our dataset contains representative information for distinguishing Poor and Good classes.

Table 5 shows the performance results of the ML classifiers for 3 different feature datasets. SVM yielded the highest performance. In Fig. 8, the accuracy values of the DL-based ML classifiers and the multi-channel 3D-CNN model are given for comparison. Although SVM achieved a better performance values than the other 5 ML classifiers, the DL classified motor impairment with the highest AUC, accuracy, F1, and especially precision values, as shown in Tables 3 & 5.

## 5. Discussion

The present study used a multi-channel 3D-CNN model to classify upper-limb motor impairment in stroke individuals. A DL technique was chosen, since other ML classifiers require complex feature analysis and because the selection of the most accurate features was a major challenge. For motor classification using DL, a dataset of both stroke patients and healthy individuals from 3 different sources was compiled. The importance of DTI maps, GM, and WM integrity for motor impairment classification was determined using a single-channel CNN. It was investigated whether ROI (motor regions + CST) or whole brain performed better when used as input in DL. The features obtained from the second FCL of DL as inputs were used for the ML classifiers to determine the best classifier for detecting upper-limb motor impairment. In the analysis of single-channel CNN, the performance values of DTI and MRI images were found to be very similar. In addition, WM, MD, and FA obtained yielded higher values, while the performance yielded by GM integrity was inferior. In the multi-channel 3D CNN, the whole-brain image analysis yielded better performance than ROIs. In addition, the highest ROC values were obtained for the 3D-CNN model with residual block, as revealed by the performance of DL and DL-based ML models.

This paper aimed to determine which type of brain variables were the most informative for classifying the level of motor impairment. For this, a separate single-channel CNN was trained and tested for each image type using DTI maps, as well as GM and WM integrity extracted from MR images. Although the performance results were very similar, the best results were obtained using WM, MD, and FA, while the worst results were obtained when using GM images. DTI maps, as well as WM and GM integrity have been shown to be potential markers for evaluating motor recovery after stroke [3,5–7,18–20]. For this reason, feature extraction was initially performed in parallel with the multi-channel 3D-CNN, where each of the DTI maps and MR images was applied to a single channel. These features were classified with both 3D-CNN and ML techniques.

Two CNN models were trained and tested in which whole-brain images and ROI images were used as input. Our performance results

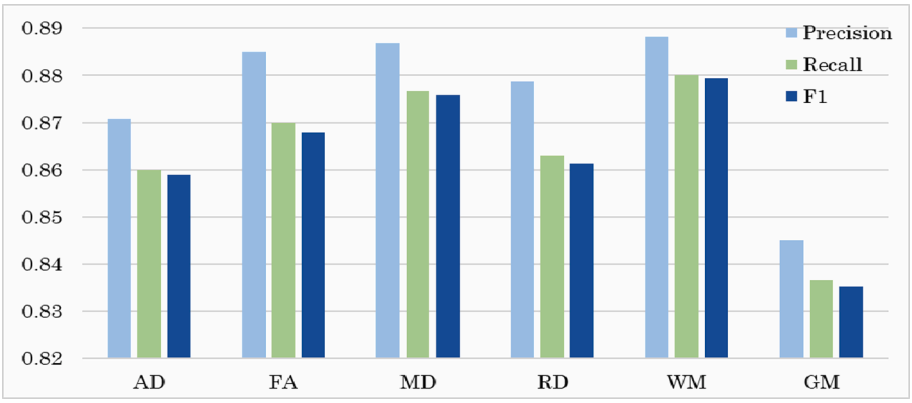


**Fig. 5.** The precision, recall, and F1 results of CNN networks trained with DTI maps, WM and GM intensity maps.

**Table 3**
Performance results of the multi-channel 3D-CNN for the whole-brain images.

| Metrics | | Poor | Good | Overall |
|---|---|---|---|---|
| Overall Mean | AUC | 0.877 [0.774 0.940] | 0.877 [0.774 0.940] | 0.877 [0.774 0.940] |
| | **Specificity** | 0.860 [0.674 0.953] | 0.895 [0.780 1.000] | 0.877 [0.727 0.977] |
| | **Accuracy** | 0.877 [0.770 0.940] | 0.877 [0.770 0.940] | 0.877 [0.770 0.940] |
| | **Precision** | 0.873 [0.741 0.955] | 0.900 [0.802 1.000] | 0.886 [0.772 0.978] |
| | **Recall** | 0.895 [0.780 1.000] | 0.860 [0.674 0.953] | 0.877 [0.727 0.977] |
| | **F1** | 0.879 [0.788 0.940] | 0.873 [0.738 0.940] | 0.876 [0.763 0.940] |
| **Mean of Maxs** | **AUC** | 0.917 [0.867 0.967] | 0.917 [0.867 0.967] | **0.917 [0.867 0.967]** |
| | **Specificity** | 0.913 [0.800 1.000] | 0.920 [0.733 1.000] | **0.917 [0.767 1.000]** |
| | **Accuracy** | 0.917 [0.867 0.967] | 0.917 [0.867 0.967] | **0.917 [0.867 0.967]** |
| | **Precision** | 0.921 [0.833 1.000] | 0.927 [0.789 1.000] | **0.924 [0.811 1.000]** |
| | **Recall** | 0.920 [0.733 1.000] | 0.913 [0.800 1.000] | **0.917 [0.767 1.000]** |
| | **F1** | 0.916 [0.846 0.966] | 0.916 [0.867 0.968] | **0.916 [0.856 0.967]** |

**Table 4**
Performance results of the multi-channel 3D-CNN for the motor region images.

| Metric | | Poor | Good | Overall |
|---|---|---|---|---|
| Overall Mean | AUC | 0.850 [0.763 0.900] | 0.850 [0.763 0.900] | 0.850 [0.763 0.900] |
| | **Specificity** | 0.820 [0.613 0.913] | 0.879 [0.793 0.967] | 0.850 [0.703 0.940] |
| | **Accuracy** | 0.850 [0.763 0.927] | 0.850 [0.763 0.927] | 0.850 [0.763 0.927] |
| | **Precision** | 0.844 [0.709 0.918] | 0.890 [0.823 0.972] | 0.867 [0.766 0.945] |
| | **Recall** | 0.879 [0.793 0.967] | 0.820 [0.613 0.913] | 0.850 [0.703 0.940] |
| | **F1** | 0.852 [0.763 0.928] | 0.839 [0.721 0.925] | 0.846 [0.742 0.927] |
| **Mean of Maxs** | **AUC** | 0.887 [0.800 0.933] | 0.887 [0.800 0.933] | 0.887 [0.800 0.933] |
| | **Specificity** | 0.800 [0.600 0.933] | 0.973 [0.933 1.000] | 0.887 [0.767 0.967] |
| | **Accuracy** | 0.887 [0.800 0.933] | 0.887 [0.800 0.933] | 0.887 [0.800 0.933] |
| | **Precision** | 0.835 [0.714 0.933] | 0.971 [0.917 1.000] | 0.903 [0.815 0.967] |
| | **Recall** | 0.973 [0.933 1.000] | 0.800 [0.600 0.933] | 0.887 [0.767 0.967] |
| | **F1** | 0.897 [0.833 0.938] | 0.873 [0.750 0.933] | 0.885 [0.792 0.935] |

were better than those reported in previous studies [23–25]. For instance, Rondina et al. [23] obtained ROI images by combining CST from MR images and the same motor and sensory regions as in the present study, which were selected using the AAL atlas. In their study, some voxels were eliminated when controlling for ROI intersections. Using the voxels in the ROI images, they classified upper-limb motor functions as Good and Poor using SVM and obtained an accuracy value of 90 %. They also trained SVM using voxels from whole-brain images and achieved an accuracy value of 80 %. Our study obtained an accuracy value of 91.7 % for the multi-channel CNN using whole-brain images and 91 % for the CNN-based SVM model. However, in their study, the dataset included images from only 30 S patients and did not include healthy controls. Therefore, the success of SVM in evaluating the motor functions for healthy controls was not available. In our study, since the Good class also included healthy controls, the corresponding accuracy (91.7 %) can thus be considered as successful for classifying healthy participants. In addition, although the study mentioned that only T1W MR images were used, CST masks were derived from DTI images obtained in 9 healthy volunteers from a previous study [63]. In our study, performance results showed that multi-channel CNN trained with whole-brain images was more successful in classifying upper-limb motor functions than the one using ROIs. The reason for this may be that some information in the ROI images was lost as a result of convolution, pooling, and activation processes used in feature mapping in the deep network. On the other hand, analyzing with whole-brain images in DL eliminated the need for manual selection and complex feature analysis.

In the literature, different ML techniques relied on various biomarkers and their different combinations to reveal the contribution of each feature to motor score estimation and classification. For instance, Hope *et al.* [64] proposed a procedure to evaluate speech impairment post-stroke using lesion information segmented from T1-weighted MR images, demographics, and behavioral data of patients in non-linear regression models. They obtained a cross-validated R squared ($R^2$) values between 0.01 and 0.84 for different biomarker combinations. In another study, Rondina *et al.* [43] decoded the motor scores of 50 chronic stroke patients using a Gaussian Process Regression (GPR) model. The correlation coefficient (R) of the GPR model was between 0.68 and 0.83 with the use of different inputs such as the whole brain, sensorimotor regions, CST, a mask of fMRI images obtained from healthy controls during handgrips, lesion location obtained from voxel-based lesion-symptom maps, and lesion-boundary mask [43]. Tozlu *et al.* [24] used five ML methods, namely elastic net, RF, artificial neural network, SVM, and classification and regression trees to predict and classify the FMA upper-limb scores using demographic, clinical, TMS-based neurophysiological, and regional dysconnectivity measurements from T1-weighted MRI inputs. The $R^2$ values of the regression models

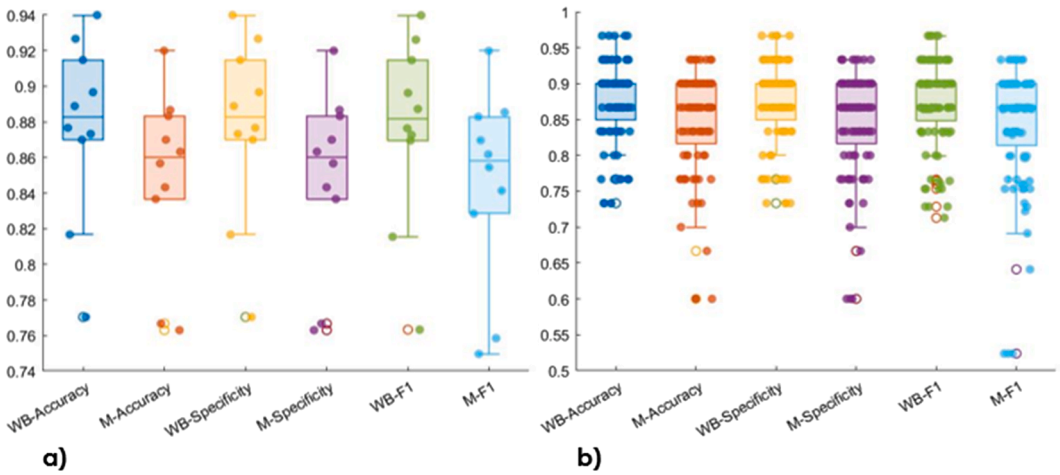


**Fig. 6.** Performance comparison of CNN Models between whole-brain (WB) and motor region (M) images, a) maximum performance results for 10 RSVs, b) performance results for 10 RSVs × 10 repetitions.

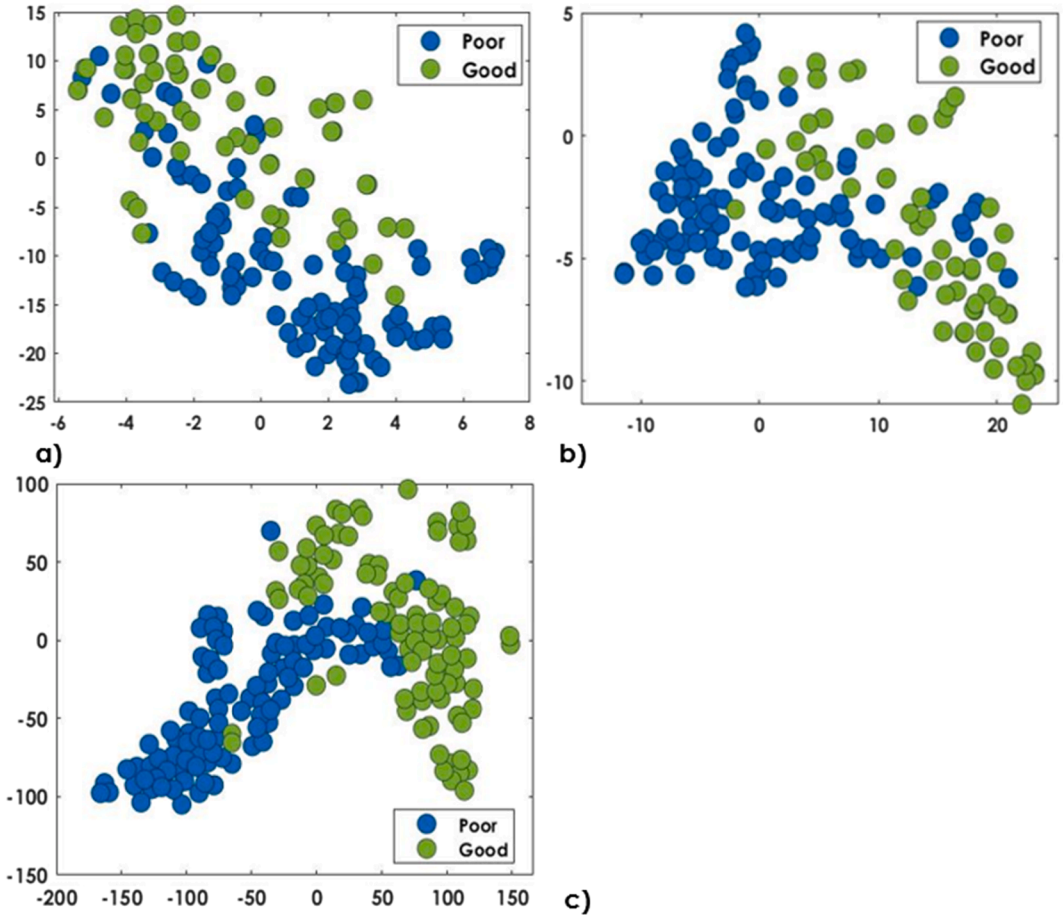


**Fig. 7.** *t*-SNE visualizations of the features obtained from the (a) flatten, (b) first, and (c) second FCLs of the 3D-CNN model.

were between 0.70 and 0.91, on the other hand, the AUC values of these classification models and logistic regression were low, between 0.50 and 0.63 [24]. On the other hand, the proposed DL-based method yielded high AUC values when using whole-brain images. In comparison with other studies where ML techniques other than DL were used, the classification performance when using whole brain information was lower [23], while feature selection and combining different biomarkers was also challenging [24,43]. Therefore, our results show that CNN can extract valuable features from whole-brain images in motor deficit classification rather than selecting specific features.

Rehme *et al.* [25] classified the hand motor impairment of 60 individuals using a two-class SVM model according to 1 of 3 groups: hand-impaired stroke patients, non-impaired stroke patients, and healthy controls. They used functional connectivity maps between the ipsilesional primary motor area (M1) region and resting-state fMRI of the whole brain, as well as voxels of DWI lesion maps as inputs in the SVM.

**Table 5**
Performance results of the ML classifiers on the features of whole-brain images.

| Classifier | Flatten | | | | FCL1 | | | | FCL2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Acc[b] | Spec[c] | F1 | AUC | Acc[b] | Spec[c] | F1 | AUC | Acc[b] | Spec[c] | F1 |
| **KNN** | 0.833 | 0.833 | 0.800 | 0.842 | 0.870 | 0.870 | 0.833 | 0.878 | 0.863 | 0.863 | 0.853 | 0.867 |
| **SVM** | 0.847 | 0.847 | 0.787 | 0.855 | 0.910 | 0.910 | 0.907 | 0.912 | 0.910 | 0.910 | 0.893 | 0.914 |
| **DT** | 0.767 | 0.767 | 0.627 | 0.796 | 0.817 | 0.817 | 0.733 | 0.832 | 0.843 | 0.843 | 0.780 | 0.854 |
| **RF** | 0.773 | 0.773 | 0.700 | 0.788 | 0.820 | 0.820 | 0.787 | 0.827 | 0.810 | 0.810 | 0.747 | 0.822 |
| **AB** | 0.830 | 0.830 | 0.780 | 0.840 | 0.843 | 0.843 | 0.767 | 0.855 | 0.883 | 0.883 | 0.820 | 0.892 |
| **NB** | 0.770 | 0.770 | 0.660 | 0.794 | 0.677 | 0.677 | 0.513 | 0.730 | 0.753 | 0.753 | 0.700 | 0.769 |

[a]FCL = Fully Connected Layer.
[b] Acc = Accuracy.
[c] Spec = Specificity.

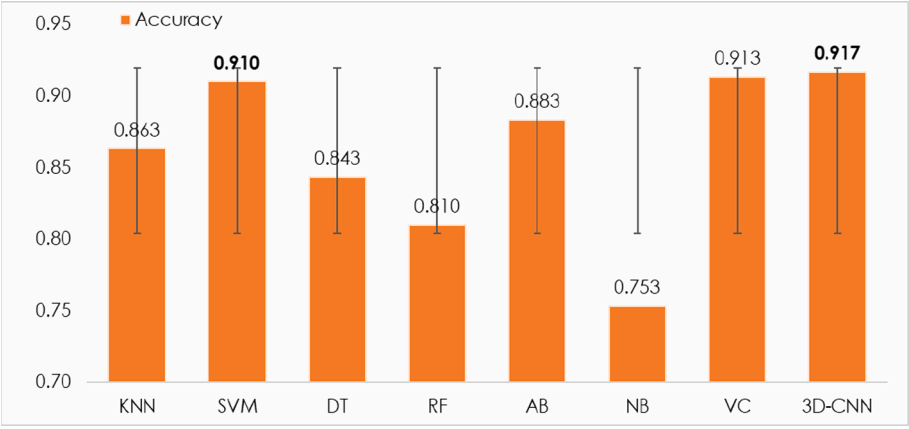


**Fig. 8.** Accuracy comparison of ML classifiers.

The classification accuracy values using functional connectivity maps were between 0.83 and 0.88 when comparing control and non-impaired subjects, and the mean accuracy was 0.85. In comparison, SVM had an accuracy value of 0.74 using the voxels of DWI lesion maps [25]. Using DL-based SVM, our study found 0.91 as the mean accuracy value for diffusivity measures derived from whole-brain DTI images, as well as GM and WM obtained from structural MR images. Although it is meaningful to use features obtained from DL layers in ML models, the multi-channel 3D-CNN with residual blocks achieved the highest accuracy.

The DL model using MRI-based images was able to predict motor recovery once the training was completed with correctly labeled classes using test scores. FMA scores were used as ground truth criteria in class labeling to prepare the data for DL training. FMA is widely used as a clinical and research tool to evaluate changes in motor impairment and motor recovery after stroke [65]. One limitation of the model is that FMA scores were obtained at one point only. Our study does not consider the motor recovery trajectory per se, including the underlying endogenous neuroplasticity events taking place post-stroke [66]. In addition, while motor impairment of very severe stroke patients is usually unpredictable using physical examination, including test scores, the structural integrity of white matter and CST-related changes can be accurately predicted using information obtained from MRI, such as lesion size and location [7]. The present study further demonstrates that there exist brain features that are associated with motor performance in individuals post-stroke. In this context, our study has shown that the proposed DL model was able to predict motor recovery using only neuroimaging data once the training was completed with correctly labeled classes using test scores. In the future, applying similar approaches to data with larger sample sizes corresponding to the acute, subacute, and chronic stages, longitudinal learning of recovery would be of interest. In particular, neuroimaging data and test scores of the motor recovery processes for the same patients followed in the clinics may result in better learning. Training DL models using a class-balanced dataset containing all stroke stages with a sufficient number of samples is a promising approach for predicting an individual's degree of motor impairment and recovery. In the case of stroke rehabilitation, it is important to consider integrating traditional physical examination with imaging features for providing a more comprehensive evaluation of a patient's progress. This could be done by harmonizing clinical tests across the cohorts being studied.

A validation period with real-world data is needed for an AI-based model to ensure its effectiveness and safe use in clinics. Evaluation metrics are crucial when assessing a predictive model in validation period and clinical settings. High sensitivity is important to ensure that all positive cases are identified. On the other hand, high specificity is important to minimize the number of false positive predictions. Hard samples, which are difficult to classify because of their high similarity to healthy individuals, increase the FN, thus threatening the recall (sensitivity) and F1-score performance. On the other hand, since the sensitivity measure is more affected by FN, a low sensitivity value during clinical practice requires improving the model with a dataset enriched with hard examples. However, this risks increasing the FP rate, since hard sample data can be confused with healthy individuals. This risk can be measured with specificity and precision metrics more sensitive to FP. As a result, in clinical practice, high sensitivity and specificity values can be achieved by observing the balance between the classes of the datasets. The balance between classes will also improve accuracy and F1-score performance.

This study achieved higher accuracy than traditional ML techniques in which feature selection is challenging. In addition, using whole-brain MRI and DTI metrics in the multi-channel 3D-CNN model has extracted richer features than ROIs. The best result obtained in this study is promising as to predict with high accuracy (92 %), but still, there is a proportion that is not well classified properly. This raises several ethical issues for the use of AI-based clinical support applications. These

applications are only as reliable as the data and algorithms used to develop them. There is always the possibility of error or bias, which could have detrimental consequences when establishing a diagnosis/prognosis or choosing a treatment. For example, a healthcare professional who follows the wrong recommendation of a DL-based system could deny the necessary treatment/care, which could adversely affect a patient's health. Before post-stroke predictive models can be implemented in clinics, they would have to be tested with the class-balanced dataset containing longitudinal data across the different phases post-stroke. Such a model would have to integrate traditional physical examination with imaging features to provide a more comprehensive evaluation of a patient's progress. This implies harmonizing clinical tests across the cohorts being studied. Continuous monitoring of the model performance and updating the dataset for each stroke phase and hard samples would also be needed. This monitoring should be carried out through collaboration between clinicians, researchers, and AI experts.

### 5.1. Limitations

There are some limitations in the present study. First, the motor impairment of stroke patients in Dataset-II and III was based on FMA scores, while a PCA score based on three different motor tests, including GS, NHPT, and BBT was used to evaluate motor deficits in Dataset-I. Since there is no matching value map between these two score groups, the motor profiles determined within each group may have adversely affected the classification performance. Second, the lesion sides of stroke patients with motor impairment and those without motor deficits were not heterogeneous. Although the images of the patients with lesions in the left hemisphere were flipped, perfect alignment could not be ensured because the brain is slightly asymmetrical due to morphological and functional differences between the hemispheres [67]. This asymmetry may constrain the performance of the computer-based analysis.

The training process for the multi-channel CNNs was performed on the CPU due to GPU memory limitation, and therefore it resulted in a long training time. The sample size used in the CNN was 154. It is expected that DL performance would be higher if more samples would be added to the dataset. Furthermore, another way to improve the performance of the DL model would be to increase the sample size using simulated DTI and MR images based on Generative Adversarial Networks (GANs). Unlike traditional data augmentation techniques (rotation, flipping, scaling, etc.), GAN has been shown to be able to learn brain structure and anomalies and produce new data that cannot be distinguished from the real ones. This would lead to the production of brain images containing lesions with a location consistent with the motor score using GAN. Thus, the effect of data augmentation techniques on the generalization ability of DL architectures in classifying motor impairment could be tested and observed.

## 6. Conclusion

The prediction of the level of motor impairment and expected recovery after stroke has received a great deal of attention in the past years. However, the prediction of upper-limb impairment based on brain features post-stroke remains challenging. This is due to the fact that combining different biomarkers, such as neuroimaging-based ROIs, as well as determining feature selection and dimension reduction methods for linear or non-linear models require trials using different combinations and algorithms. The present study aimed to generate a classifier based on a DL approach to tackle this challenge. To our knowledge, it is the first study in the literature that has used DL or DL-based ML models on a dataset of this size in the context of stroke. The proposed approach developed a multi-channel 3D-CNN model with a residual block that evaluates a combination of multimodal feature sources without losing important but low-gradient information.

The single-channel 3D-CNN trained with WM, MD, and FA achieved slightly better ROC values than the other diffusivity maps and GM images. The results suggest that our multi-channel CNN is able to extract important features from whole-brain images which include rich information, rather than selecting features from ROIs such as lesions or motor regions. The proposed 3D-CNN model using residual blocks successfully identified motor deficits with an accuracy of 0.92. Our approach also has the advantage of not requiring feature analysis and manual region selection, unlike other ML techniques. It has the potential to be easily implemented in clinical settings where big datasets are available. It could be used to better predict motor recovery early on after stroke, which could ultimately promote the use of targeted rehabilitation interventions for maximizing the functional independence of these individuals.

## CRediT authorship contribution statement

**Rukiye Karakis:** Conceptualization, Methodology, Data curation, Software. **Kali Gurkahraman:** Data curation, Software. **Georgios D. Mitsis:** Supervision, Writing – review & editing. **Marie-Hélène Boudrias:** Writing – review & editing, Project administration, Supervision.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank Dr. Lara A. Boyd from the University of British Columbia, Vancouver, BC, Canada, Dr. Adrian G. Guggisberg from the University of Geneva, Switzerland, and Dr. Nick Ward from the UCL Institute of Neurology, United Kingdom, England, for sharing the data included in this article. We also like to thank Arna Ghosh, Drs. Danilo Bzdok and Benjamin De Leener for helpful comments on the use of DL. Data from the Boyd lab were collected using funds from the Canadian Institutes of Health Research (PI: Boyd, MOP-130269). This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) 2219 project program [RK] and the Canadian Foundation for Innovation grant number 34277 [MHB]. At the time of the study, MHB held a Fonds de recherche du Québec–Santé (FRQ-S) Research Scholar Award.

## References

[1] World Health Organization, Cerebrovascular Disorders (Offset Publications). Health Organization, ISBN 978-92-4-170043-6. OCLC 4757533, World, Geneva, 1978.
[2] Y. Zhang, A.M. Chapman, M. Plested, D. Jackson, F. Purroy, The incidence, prevalence, and mortality of stroke in France, Germany, Italy, Spain, the UK, and the US: a literature review, Stroke Res. Treatment. (2012), 436125.
[3] L.M. Moura, R. Luccas, J. de Paiva, E. Amaro, A. Jr Leemans, C. Leite, M. Otaduy, A.B. Conforto, Diffusion tensor imaging biomarkers to predict motor outcomes in stroke: a narrative review, Front. Neurol. 10 (2019) 445.
[4] S. Larivière, N.S. Ward, M.H. Boudrias, Disrupted functional network integrity and flexibility after stroke: relation to motor impairments, NeuroImage: Clinical. 19 (2018) 883–891.
[5] L.A. Boyd, K.S. Hayward, N.S. Ward, C.M. Stinear, C. Rosso, R.J. Fisher, A. R. Carter, A.P. Leff, D.A. Copland, L.M. Carey, L.G. Cohen, D.M. Basso, J. M. Maguire, S.C. Cramer, Biomarkers of stroke recovery: consensus-based core recommendations from the stroke recovery and rehabilitation roundtable, Int. J. Stroke 12 (5) (2017) 480–493.
[6] M.K.I. Zolkefley, Y.M.S. Fiwana, H.Z.M. Hatta, C. Rowbin, C.M.N.C.M. Nassir, M. H. Hanafi, M.S. Abdullah, M. Mustapha, An overview of fractional anisotropy as a reliable quantitative measurement for the corticospinal tract (CST) integrity in correlation with a Fugl-Meyer assessment in stroke rehabilitation, J Phys Ther Sci. 33 (1) (2021) 75–83.
[7] J. Puig, G. Blasco, G. Schlaug, C.M. Stinear, P. Daunis-I-Estadella, C. Biarnes, J. Figueras, J. Serena, M. Hernández-Pérez, A. Alberich-Bayarri, M. Castellanos, D. S. Liebeskind, A.M. Demchuk, B.K. Menon, G. Thomalla, K. Nael, M. Wintermark, S. Pedraza, Diffusion tensor imaging as a prognostic biomarker for motor recovery and rehabilitation after stroke, Neuroradiology 59 (4) (2017) 343–351.
[8] D.J. Werring, A.T. Toosy, C.A. Clark, G.J. Parker, G.J. Barker, D.H. Miller, A.J. Thompson Diffusion tensor imaging can detect and quantify corticospinal tract degeneration after stroke, J. Neurol. Neurosurg. Psychiatry 69 (2) (2000) 269–272.

[9] M.H. Milot, S.C. Cramer, Biomarkers of recovery after stroke, Curr. Opin. Neurol. 21 (6) (2008) 654–659.
[10] C. Stinear, Prediction of recovery of motor function after stroke, Lancet Neurol. 9 (2010) 1228–1232.
[11] C.H. Park, N. Kou, M.H. Boudrias, E.D. Playford, N.S. Ward, Assessing a standardised approach to measuring corticospinal integrity after stroke with DTI, Neuro Image: Clinical. 2 (2013) 521–533.
[12] C.M. Stinear, P.A. Barber, M. Petoe, S. Anwar, W.D. Byblow, The PREP algorithm predicts the potential for upper limb recovery after stroke, Brain 135 (2012) 2527–2535.
[13] C.M. Stinear, W.D. Byblow, S.J. Ackerley, M.C. Smith, W.M. Borges, A.P. Barber, PREP2: a biomarker-based algorithm for predicting upper limb function after stroke, Ann. Clin. Transl. Neurol. 4 (11) (2017) 811–820.
[14] A.L. Alexander, J.E. Lee, M. Lazar, A.S. Field, Diffusion tensor imaging of the brain, Neurotherapeutics 4 (3) (2007) 316–329.
[15] S.N. Niogi, P. Mukherjee, Diffusion tensor imaging of mild traumatic brain injury, J. Head Trauma Rehabilitation. 25 (2010) 241–255.
[16] M. O'Sullivan, Imaging small vessel disease: lesion topography, networks, and cognitive deficits investigated with MRI, Stroke 41 (10 Suppl) (2010) S154–S158.
[17] Z. Chen, P. Ni, J. Zhang, Y. Ye, H. Xiao, G. Qian, S. Xu, J. Wang, X. Yang, J. Chen, B. Zhang, Y. Zeng, Evaluating ischemic stroke with diffusion tensor imaging, Neurol. Res. 30 (7) (2008) 720–726.
[18] Q. Diao, J. Liu, C. Wang, C. Cao, J. Guo, T. Han, J. Cheng, X. Zhang, C. Yu, Gray matter volume changes in chronic subcortical stroke: a cross-sectional study, NeuroImage: Clinical. 14 (2017) 679–684.
[19] C. Dang, G. Liu, S. Xing, C. Xie, K. Peng, C. Li, J. Li, J. Zhang, L. Chen, Z. Pei, J. Zeng, Longitudinal cortical volume changes correlate with motor recovery in patients after acute local subcortical infarction, Stroke 44 (2013) 2795–2801.
[20] M. Yang, Y.-R. Yang, H.-J. Li, X.-S. Lu, Y.-M. Shi, B. Liu, H.-J. Chen, G.-J. Teng, R. Chen, E.H. Herskovits, Combining diffusion tensor imaging and gray matter volumetry to investigate motor functioning in chronic stroke, PLoS One 10 (5) (2015) e0125038.
[21] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, J. Wang, Applications of deep learning to MRI images: a survey, Big Data Mining and Anal. 1 (1) (2018) 1–18.
[22] G. Zaharchuk, E. Gong, M. Wintermark, D. Rubin, C.P. Langlotz, Deep Learning in neuroradiology, AJNR Am. J. Neuroradiol. 39 (10) (2018) 1776–1784.
[23] J.M. Rondina, C.H. Park, N.S. Ward, Brain regions important for recovery after severe post-stroke upper limb paresis, J. Neurol. Neurosurg. Psychiatry 88 (9) (2017) 737–743.
[24] C. Tozlu, D. Edwards, A. Boes, D. Labar, K.Z. Tsagaris, J. Silverstein, A. Kuceyeski, Machine learning methods predict individual upper-limb motor impairment following therapy in chronic stroke, Neurorehabil. Neural Repair 34 (5) (2020) 428–439.
[25] A.K. Rehme, L.J. Volz, D.-L. Feis, I. Bomilcar-Focke, T. Liebig, S.B. Eickhoff, G. R. Fink, C. Grefkes, Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques, Cereb. Cortex 25 (9) (2015) 3046–3056.
[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012, pp. 1097-1105.
[27] Y. LeCun, Y. Bengio, G.E. Hinton, Deep Learning, Nature 521 (2015) 436–444.
[28] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recogn. 77 (2018) 354–377.
[29] S.P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, B. Gulyás, 3D deep learning on medical images: a review, Sensors 20 (18) (2020) 5097.
[30] A. Tiwari, S. Srivastava, M. Pant, Brain tumor segmentation and classification from magnetic resonance images: review of selected methods from, to 2019, Pattern Recogn. Lett. 131 (2020) (2014) 244–260.
[31] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, G. Catheline, 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies. 2018, arXiv:1801.05968.
[32] K. Gurkahraman, R. Karakis, Brain tumors classification with deep learning using data augmentation, J. Fac. Eng. Archit. Gazi Univ. 36 (2) (2021) 997–1011.
[33] E.-J. Lee, Y.-H. Kim, N. Kim, D.-W. Kang, Deep into the brain: artificial intelligence in stroke imaging, J. Stroke. 19 (3) (2017) 277–285.
[34] R. Karthik, R. Menaka, A. Johnson, S. Anand, Neuroimaging and deep learning for brain stroke detection- a review of recent advancements and future prospects, Comput. Methods Programs Biomed. 197 (2020), 105728.
[35] A. Nielsen, M.B. Hansen, A. Tietze, K. Mouridsen, Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning, Stroke 49 (6) (2018) 1394–1401.
[36] J. Heo, J.G. Yoon, H. Park, Y.D. Kim, H.S. Nam, J.H. Heo, Machine learning-based model for prediction of outcomes in acute stroke, Stroke 50 (5) (2019) 1263–1265.
[37] V. Mathiowetz, K. Weber, G. Volland, N. Kashman, Reliability and validity of grip and pinch strength evaluations, J. Hand Surg. 9 (2) (1984) 222–226.
[38] V. Mathiowetz, K. Weber, N. Kashman, G. Volland, Adult norms for the nine hole peg test of finger dexterity, The Occupational Therapy J. Res. 5 (1) (1985) 24–38.
[39] V. Mathiowetz, G. Volland, N. Kashman, K. Weber, Adult norms for the box and block test of manual dexterity, Am. J. Occup. Ther. 39 (1985) 386–391.
[40] K.S. Hayward, J.L. Neva, C.S. Mang, S. Peters, K.P. Wadden, J.K. Ferris, L.A. Boyd, Interhemispheric pathways are important for motor outcome in individuals with chronic and severe upper limb impairment post stroke, Neural Plast. (2017), 4281532.
[41] A.R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, S. Steglind, The post-stroke hemi-237 plegic patient a method for evaluation of physical performance, Scand J. Rehabil. Med. 7 (1) (1975) 13–31.
[42] A.G. Guggisberg, P. Nicolo, L.G. Cohen, A. Schnider, E.R. Buch, Longitudinal structural and functional differences between proportional and poor motor recovery after stroke, Neurorehabil. Neural Repair 31 (12) (2017) 1029–1041.
[43] J.M. Rondina, M. Filippone, M. Girolami, N.S. Ward, Decoding post-stroke motor function from structural brain imaging, NeuroImage Clinical. 12 (2016) 372–380.
[44] E.J. Woytowicz, J.C. Rietschel, R.N. Goodman, S.S. Conroy, J.D. Sorkin, J. Whitall, S. McCombe Waller, Determining levels of upper extremity movement impairment by applying a cluster analysis to the Fugl-Meyer assessment of the upper extremity in chronic stroke, Arch. Phys. Med. Rehabil. 98 (3) (2017) 456–462.
[45] N. Hijikata, M. Kawakami, R. Ishii, K. Tsuzuki, T. Nakamura, K. Okuyama, M. Liu, Item difficulty of Fugl-Meyer assessment for upper extremity in persons with chronic stroke with moderate-to-severe upper limb impairment, Front. Neurol. 11 (2020), 577855.
[46] L.J. O'Donnell, C.-F. Westin, An introduction to diffusion tensor image analysis, Neurosurg. Clin. N. Am. 22 (2) (2011), 185–viii.
[47] J.M. Soares, P. Marques, V. Alves, N. Sousa, A hitchhiker's guide to diffusion tensor imaging, Front Neurosci. 7 (2013) 31.
[48] A. Leemans, D.K. Jones, The B-matrix must be rotated when correcting for subject motion in DTI data, Magn. Reson. Med. 61 (2009) 1336–1349.
[49] Image and video denoising by sparse 3D transform-domain collaborative filtering. http://www.cs.tut.fi/~foi/GCF-BM3D/index.html#ref_people. 2022 (accessed 10 Nov 2022).
[50] Y. Mäkinen, L. Azzari, A. Foi, Collaborative filtering of correlated noise: exact transform-domain variance for improved shrinkage and patch matching, IEEE Trans. Image Process. 29 (2020) 8339–8354.
[51] O. Hauk, I. Johnsrude, F. Pulvermüller, Somatotopic representation of action words in human motor and premotor cortex, Neuron 41 (2004) 301–307.
[52] Y. Cao, L. D'Olhaberriague, E.M. Vikingstad, S.R. Levine, K.M. Welch, Pilot study of functional MRI to assess cerebral activation of motor function after poststroke hemiparesis, Stroke 29 (1998) 112–122.
[53] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, Neuroimage 15 (1) (2002) 273–289.
[54] Q. Dou, H. Chen, L.Q. Yu, L. Zhao, J. Qin, D.F. Wang, V.C. Mok, L. Shi, P.A. Heng, Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks, IEEE Trans. Med. Imaging. 35 (5) (2016) 1182–1195.
[55] F. Behrad, M.S. Abadeh, An overview of deep learning methods for multimodal medical data mining, Expert Syst. Appl. (2022), 117006.
[56] D. Nie, J. Lu, H. Zhang, E. Adeli, J. Wang, Z. Yu, D. Shen, et al., Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages, Sci. Rep. 9 (1) (2019) 1103.
[57] S. Bianco, R. Cadene, L. Celona, P. Napoletano, Benchmark analysis of representative deep neural network architectures, IEEE Access 6 (2018) 64270–64277.
[58] S. Suganyadevi, V. Seethalakshmi, K. Balasamy, A review on deep learning in medical image analysis, Int. J. Multimedia Information Retrieval 11 (1) (2022) 19–38.
[59] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[60] E. Montagnon, M. Cerny, A. Cadrin-Chênevert, V. Hamilton, T. Derennes, A. Ilinca, A. Tang, Deep learning workflow in radiology: a primer, Insights into imaging 11 (2020) 1–15.
[61] M. Rana, M. Bhushan, Machine learning and deep learning approach for medical image analysis: diagnosis to detection, Multimed. Tools Appl. 1–39 (2022), https://doi.org/10.1007/s11042-022-14305-w.
[62] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 861–874.
[63] R. Schulz, C.H. Park, M.H. Boudrias, C. Gerloff, F.C. Hummel, N.S. Ward, Assessing the integrity of corticospinal pathways from primary and secondary cortical motor areas after stroke, Stroke 43 (8) (2012) 2248–2251.
[64] T.M. Hope, M.L. Seghier, A.P. Leff, C.J. Price, Predicting outcome and recovery after stroke with lesions extracted from MRI images, NeuroImage Clin. 2 (2013) 424–433.
[65] D.J. Gladstone, C.J. Danells, S.E. Black, The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties, Neurorehabil. Neural Repair 16 (3) (2002) 232–240.
[66] J. Bernhardt, K.S. Hayward, G. Kwakkel, N.S. Ward, S.L. Wolf, K. Borschmann, S. C. Cramer, Agreed definitions and a shared vision for new standards in stroke recovery research: the stroke recovery and rehabilitation roundtable taskforce, Neurorehabil. Neural Repair 31 (9) (2017) 793–799.
[67] S.X. Liu, Symmetry and asymmetry analysis and its implications to computer-aided diagnosis: A review of the literature.