Federated Deep Learning for the Diagnosis of Cerebellar Ataxia: Privacy Preservation and Auto-Crafted Feature Extractor

Thang Ngo, *Student Member, IEEE*, Dinh C. Nguyen, *Student Member, IEEE*, Pubudu N. Pathirana, *Senior Member, IEEE*, Louise A. Corben, Martin B. Delatycki, Malcolm Horne, David J. Szmulewicz, and Melissa Roberts

***Abstract*— Cerebellar ataxia (CA) is concerned with the incoordination of movement caused by cerebellar dysfunction. Movements of the eyes, speech, trunk, and limbs are affected. Conventional machine learning approaches utilizing centralised databases have been used to objectively diagnose and quantify the severity of CA. Although these approaches achieved high accuracy, large scale deployment will require large clinics and raises privacy concerns. In this study, we propose an image transformation-based approach to leverage the advantages of state-of-the-art deep learning with federated learning in diagnosing CA. We use motion capture sensors during the performance of a standard neurological balance test obtained from four geographically separated clinics. The recurrence plot, melspectrogram, and poincaré plot are three transformation techniques explored. Experimental results indicate that the recurrence plot yields the highest validation accuracy (86.69%) with MobileNetV2 model in diagnosing CA. The proposed scheme provides a practical solution with high diagnosis accuracy, removing the need for feature engineering and preserving data privacy for a large-scale deployment.**

***Index Terms*— Deep learning (DL), transfer learning, federated learning (FL), cerebellar ataxia (CA).**

Manuscript received October 16, 2021; revised March 2, 2022; accepted March 16, 2022. Date of publication March 22, 2022; date of current version March 30, 2022. This work was supported in part by the Florey Institute of Neuroscience under Grant GNT1101304 and in part by the National Health and Medical Research Council (NHMRC) of Australia under Grant APP1129595. *(Corresponding author: Thang Ngo.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted in part by the Human Research and Ethics Committee, Royal Victorian Eye and Ear Hospital, East Melbourne, VIC, Australia, under Approval No. 11/994H/16; in part by the Monash Health Human Research Ethics Committee under Approval No. HREC/18/MonH/418; and in part by the Human Research and Ethics Committee, Deakin University, Australia, under Aprroval No. STEC-02-2019-RALLAPALLE; and performed in line with the Declaration of Helsinki.

Thang Ngo and Pubudu N. Pathirana are with the School of Engineering, Deakin University, Waurn Ponds, VIC 3216, Australia (e-mail: tdngo@deakin.edu.au; pubudu.pathirana@deakin.edu.au).

Dinh C. Nguyen is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: nguye772@purdue.edu; cdnguyen@deakin.edu.au).

Louise A. Corben and Martin B. Delatycki are with the Murdoch Children's Research Institute, Parkville, VIC 3052, Australia (e-mail: louise.corben@mcri.edu.au; martin.delatycki@vcgs.org.au).

Malcolm Horne is with the Florey Institute of Neuroscience and Mental Health, Parkville, VIC 3052, Australia (e-mail: malcolm.horne@florey.edu.au).

David J. Szmulewicz is with the Florey Institute of Neuroscience and Mental Health, Parkville, VIC 3052, Australia, also with the Balance Disorders and Ataxia Service, Royal Victorian Eye and Ear Hospital (RVEEH), East Melbourne, VIC 3002, Australia, and also with the Cerebellar Ataxia Clinic, Alfred Hospital, Prahran, VIC 3004, Australia (e-mail: dsz@me.com).

Melissa Roberts is with the Physiotherapy Department, Monash Health, Clayton, VIC 3168, Australia (e-mail: melissa.roberts@monashhealth.org).

Digital Object Identifier 10.1109/TNSRE.2022.3161272

I. Introduction

CEREBELLAR Ataxia (CA) is the un-coordinated movement resulting from impaired function of the cerebellum. The impaired movements are usually assessed according the following body regions: eye movements, speech, axial function (balance and gait), appendicular function (upper and lower limbs). Ataxia is diagnosed and assessed by a clinician recognising the characteristic uncoordinated movements, often by asking the subject to perform specific tasks to accentuate the incoordination. The clinician's experience and the inherent subjectiveness of the human decision-making procedure impact significantly on this assessment method.

Three fundamental problems in assessing ataxia are (i) classifying people as ataxics or non-ataxic controls; (ii) assessing severity with a regression model or classify the severity into sub-groups of low, moderate, and severe; and, (iii) classifying subjects ataxia phenotype to facilitate more specific therapeutic and rehabilitation programs. The conventional machine learning (ML) approaches involve a broad range of motion sensors to capture ataxic movements. Raw data may be processed with bias removal and signal filtering (e.g. low-pass, bandpass filter) to remove noise from the internal and external environment of the device. However, ataxic movements are considered ad-hoc movements by which frequency bands that manifest ataxia are an open research question. The standalone features or the combined ones that are captured from the sensory information is fed into ML models in achieving the aforementioned classifications.

Deep learning (DL) is a subdomain of ML that has witnessed sensational advancement recently, owing to increased computer power and the availability of big datasets [2], [3]. Normally, developing a ML scheme requires domain understanding and often embedded with specific human to construct feature extractors transforming raw data into appropriate representative forms. This process, often called hand-crafted feature extraction, is inherently onerous and tentative albeit focused on obtaining quality features to train ML models. On the

contrary, DL is a representative learning in which the model itself extracts the features needed for pattern recognition from the raw input data. DL, called auto-crafted feature extraction, stacks multiple computational layers sequentially. Each layer is composed of a large number of primitive, nonlinear operations, and the representation of each layer is sent into the next layer, which transforms it into a more abstract representation.

In the health care sector, institutions are reluctant to share clinical datasets containing confidential information without stringent safeguards. Institutional privacy laws, guidelines and rules for keeping and transmitting personal health data are enforced by the European General Data Protection Regulation (GDPR) and the United States Health Insurance Portability and Accountability Act (HIPAA) [4]. Under the constraints of these restrictions, DL models trained on one local dataset may not be generalizable and function poorly when reused on another dataset. This is often the case in less diverse medical datasets with limited samples. Federated learning (FL) has emerged as a promising solution that protects the privacy of patient data while achieving DL by distributed AI to combine local (individual-level) and global (group-level) models [5], [6]. This learning paradigm is achieved by performing collaborative AI training on data from multiple sources that have been aggregated in a cloud server (for example). In this manner, each participating clinic would train the global DL model using their local CA dataset and only exchange model parameters such as gradients. This methodology reduces the cost of data transmission while protecting privacy without any requirement to sharing actual data containing sensitive patient information.

Recently, FL and DL have been applied in the health care sector and have achieved high accuracies of 89% on human activity recognition [7]–[9]. FL has been utilized for a cloud-edge based framework for personalized in-home health monitoring [10]. A review on FL has considered the challenges and explores its potential to offer a solution for the future expansion of digital health systems [11]. In medical diagnostic processes, FL and DL have been combined to detect COVID-19 using image-based data [12]. Beyond FL, Swarm Learning has recently been introduced as a decentralized ML approach that unites edge computing, blockchain-based peer-to-peer networking and coordination [13]. With respect to measurement of CA, data is often captured from either neuroimaging or sensor systems. With a 13.75% percent error rate, magnetic resonance imaging (MRI) of the brain has been utilised to distinguish between three CA genotypes [Spinocerebellar ataxia type 2 (SCA2), SCA6, and Ataxia-telangiectasia (AT)] [14].

There are a broad range of approaches to measure ataxia using sensors. For example, movement of the eye's iris was captured by a mobile phone camera to diagnose CA with 84% sensitivity and 77% specificity [15]. Ataxic speech has been analysed by combining perceptual and acoustic measures to distinguish between subjects with Friedreich Ataxia (FRDA) and control participants with more than 80% accuracy [16]. Assessment of upper limb ataxia has been achieved using data from a depth sensing Microsoft Kinect© camera based system to quantify the instability of the index finger [17], electromyography (EMG) to quantify the motor functions of individuals with CA [18], or Q-motor system to evaluate a lift, finger tapping, and pronate/supinate tasks [19]. Researchers also used prism-equipped goggles to assess motor adaptation in ataxia [20], or wrist worn IMU motion sensor to separate individuals with ataxia and controls who performed finger tapping and fast alternative hand movements [21]. In the balance domain, a cloud-based ML implementation operating on data from IMU based wearable sensors were used to quantify the severity of truncal ataxia [1], [22]. Kinect© cameras [23], kinematic sensors [24]–[27], pressure-sensitive walkway [28], infrared camera [29], and S-band sensing [30] framework were employed to differentiate between the gait of individuals with and without ataxia. The systems described above modeled motor tasks assessed using ataxia clinical scales. Sensors have also been placed in common devices (spoon [31], cup [32], and mouse [33]) to measure kinematic parameters that are then used to produce measurements that distinguish control subjects and individuals with ataxia.

Most studies in CA have used traditional ML with hand-crafted feature extraction. Out of more than 100 reported signal processing features [34], the performance of the ML model depends heavily on quantity and quality of selected features. DL applications in the CA field are limited despite outstanding achievements in other fields. Work by Yang *et al.* is the only research to use DL with neuroimaging to classify phenotype [14]. This is most likely due to the restrictions arising from the large datasets required with imaging such as clinical MRI scans. To our knowledge, we are the first group applying DL in CA diagnosis with a sensor system. We propose a DL scheme to reduce workload, combined with FL to prevent data leakage. We evaluate three different image transformation techniques to use with DL and employ transfer learning to overcome the problem of limited datasets. Indeed, our proposed solution is not exclusive to time series data and can leverage the power of DL in many other systems. The proposed scheme would provide a practical solution to enhance performance while protecting data privacy. To this end, our contributions are summarised as:

1) suggesting a DL approach in CA with image-based transformations and transfer learning to tackle limited data resources;
2) applying and testing performance of FL in a practical scenario with four participating clinics;
3) proposing a modified recurrence plot to reduce stored information by half, resulting in decreased processing time.

The structure of the paper is as follows: Section II outlines data demographic, sensor platform and data collection protocol. Next, we described signal processing, image transformation techniques, and statistical analyses in Section III; followed by an introduction to DL and the proposed federated scheme. Experiment results are given in Section IV and Section V provides concluding remarks.

II. Material

A. Participants

Datasets in this study were provided by four clinics, located in three Australian states: Victoria, New South Wales and

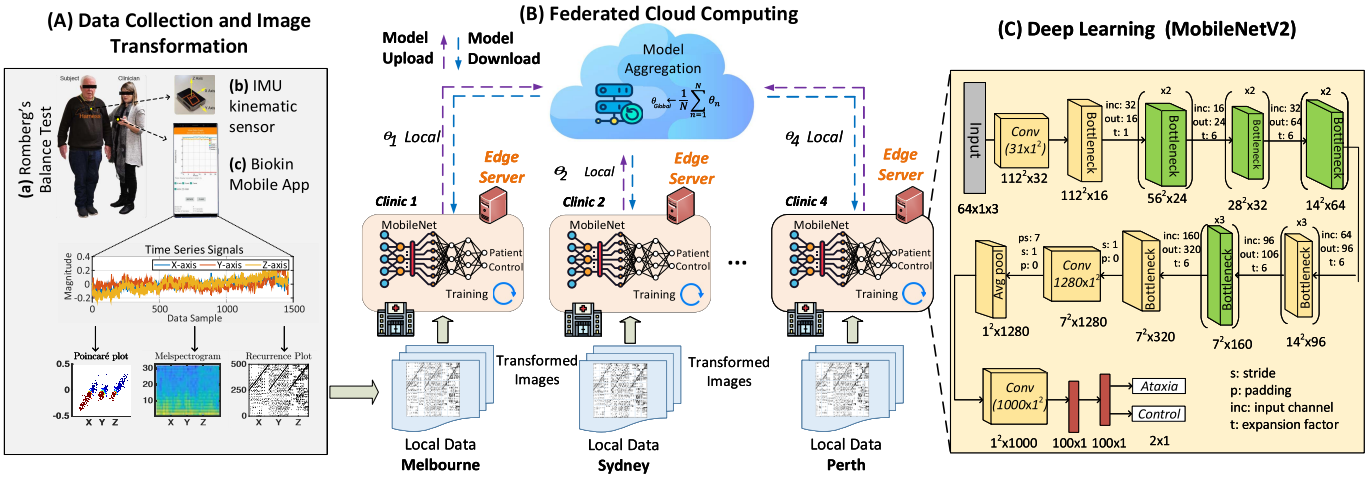


Fig. 1. Federated Deep Learning scheme to diagnose CA with Inertial measurement unit (IMU) sensor. **(A)** Data collection protocol and image transformation techniques including (a) Romberg's balance point-of-care test with an individual with ataxia, (b) a Biokin^{TM} kinematic sensor [1], and (c) Mobile application as a user interface. **(B)** Federated deep learning cloud-based implementation to analyse motion data. **(C)** Deep learning architecture.

TABLE I
DEMOGRAPHIC DATA OF PARTICIPANTS

		Controls	Individuals with Ataxia
Dataset 1	Data source	Melbourne (St. Vincent's Hospital)	
	Number of subjects (people)	24	62
	Ages, mean ± SD (years)	50.4 ± 19.9	61.0 ± 11.7
	Gender (M/F)	6/18	36/26
	Dominant limb (left/right)	6/18	11/51
	Phenotype: CA/(CABV + SS) /CABV/unknown	not applicable	25/10/11/16
Dataset 2	Data source	Supplementary Data 1	Sydney
	Number of subjects (people)	18	16
	Ages, mean ± SD (years)	35.39 ± 10.0	47.2 ± 15.9
	Gender (M/F)	7/11	6/10
	Dominant limb (left/right)	2/16	-
	Phenotype: SCA/FRDA /ANO10/unknown	not applicable	11/3/1/1
Dataset 3	Data source	Supplementary Data 2	Perth
	Number of subjects (people)	11	12
	Ages, mean ± SD (years)	33.72 ± 9.67	47.6 ± 20.3
	Gender (M/F)	6/5	6/6
	Dominant limb (left/right)	0/11	-
	Phenotype: SCA/FRDA SPG7/CANVAS/unknown	not applicable	4/2/1/2/3
Dataset 4	Data source	Supplementary Data 3	Melbourne (Monash Health)
	Number of subjects (people)	16	13
	Ages, mean ± SD (years)	34.5 ± 9.64	45.0 ± 17.5
	Gender (M/F)	11/5	10/3
	Dominant limb (left/right)	1/15	-
	Phenotype: SCA/FRDA/AOA2	not applicable	5/4/4

SD denotes standard deviation.

Western Australia. Demographic information for the data sets are detailed in Table I. Supplementary datasets 1, 2, 3 of controls were collected from Deakin University trial and matched to participants at each site. Ethics approval was granted by the Human Research and Ethics Committee, Royal Victorian Eye and Ear Hospital, East Melbourne, Australia (HREC Number: 11/994H/16). It was also approved by the Monash Health Human Research Ethics Committee (HREC Number: HREC/18/MonH/418) and the Human Research and Ethics Committee, Deakin University, Australia (HREC Reference Number: STEC-02-2019-RALLAPALLE). All participants provided informed consent as per the Declaration of Helsinki.

B. Experimental Protocol

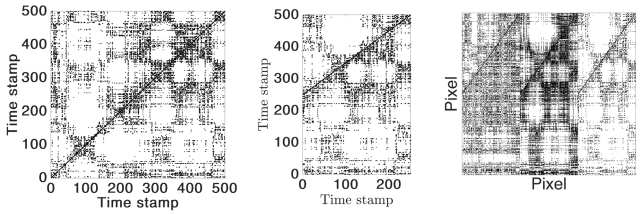
Fig. 1A depicts an individual undertaking the Romberg's balance bedside test. Two time synchronized IMU sensors were attached in the midline to the back at the level of the fourth thoracic vertebra and the chest at the manubriosternal joint. The sensor placement is referenced from work of Ghislieri *et al.* [22] and our previous research [1]. The X-, Y-, and Z-axes of the sensor are aligned with the Medial-Lateral, Superior-Inferior, and Anterior-Posterior axes, respectively. Subjects were instructed to stand in a natural standing stance with their arms besides their bodies and maintain balance. The IMU sensor recorded data for 30 seconds with the eyes open and 30 seconds with the eyes closed. The subject repeated the protocol with their feet together and apart. The IMU accelerometer sensor sampling rate set to 50 Hertz compensating for low frequency human movements and the number of samples required for image transforming techniques. Ideally each component lasted 30 seconds; however, the task could be stopped at any time if there was a risk of falling.

III. METHOD

The schematic depicted in Fig. 1B, C outlines the functional architecture in two principle aspects: FL(III-E.1) and DL (III-D).

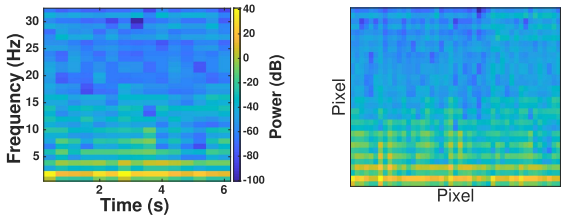
A. Signal Pre-Processing

Interpolating and sliding a fixed-sized window with two second overlapping was used to pre-process the raw data. As recurrence plots are significantly reliable when the data length is longer than 1000 samples [35], linear interpolation was applied in case the collected accelerometer signal was less than 1000 samples. To accommodate difficulty in recording



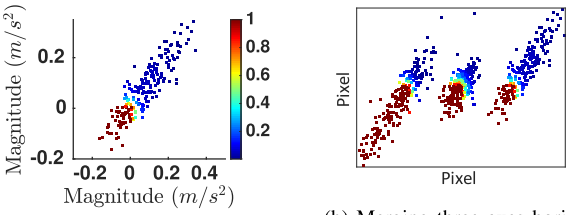
(a) Recurrence Plot, Anterior-Posterior axis. (b) Cut and Rotated Recurrence Plot. (c) Merging three axes horizontally.

Fig. 2. Recurrence plot pattern matrix. (a) Original RP. (b) Modified RP. (c) The final RP image to train the models by combining three axes.



(a) Melspectrogram, Anterior-Posterior axis. (b) Input image to train DL model.

Fig. 3. Melspectrogram pattern matrix, accelerometer, Anterior-Posterior axis. (a) Original melspectrogram. (b) Merging three melspectrograms horizontally to create input image training DL models.



(a) Poincaré plot, Anterior-Posterior axis. (b) Merging three axes horizontally.

Fig. 4. Poincaré Plot image transformation. (a) Poincaré Plot of one representative individual. (b) Three axes stack horizontally as input image to train and test DL models.

individual movements overlaid unintended movements, a sliding window of 500 samples or 10 seconds were employed to identify the most reliable data frame.

B. Image Transformation Techniques

1) Recurrence Plot: The recurrence plot (RP) was a graphical representation visualizing the stability within a signal [1]. To construct the RP of a signal $x^0(n)$ with $n \in [1, \ldots, N]$ and N is the total data sample, we first define the time delay $R^{\delta,\tau}(m)$ by

$$R^{\delta,\tau}(m) = \begin{bmatrix} x^0(m) \\ x^1(m+\tau) \\ \vdots \\ x^\delta(m+\delta\tau) \end{bmatrix}, \quad 1 \le m \le N - \delta\tau \tag{1}$$

A delayed version $x^\delta(m)$ of $x^0(n)$ is constructed with $x^\delta(m+\delta\tau) = x^0(n)$ for $n = m + \delta\tau$. Here, δ and τ are pre-defined dimension and time delay parameters, respectively. The next step is to measure Euclidean distances $D(m,h) = \|R^{\delta,\tau}(m) - R^{\delta,\tau}(h)\|$ between pairwise time delay signals $R^{\delta,\tau}(\cdot)$. RP is a square matrix of size $(N-\delta\tau)$ with element at the crossing of a column m and a row h is equal to one (black point) if $D(m,h) < \epsilon$, and zero (white point) if $D(m,h) > \epsilon$, as indicated by

$$RP_{m,h} = \Theta[\epsilon - D(m,h)] \tag{2}$$

where $\Theta(\cdot)$ is the Heaviside function and ϵ is a predefined threshold. Fig. 2 showed the RP corresponding to a typical CA individual. As the original RP contained symmetrically redundant information across the diagonal line, we removed the upper half triangle (Fig. 2a) and rotated a left portion of the lower triangle to reduce duplicative information. The image input (Fig. 2b) halved in size leaded to lesser complex DL structures and faster training times.

2) Melspectrogram: Considering the IMU signal as non periodic signal over data collection time, melspectrogram used short-time Fourier transform (STFT) to analyze the frequency components of the signal. Specifically, the STFT was computed on overlapping windowed segments of the signal and converted to the mel scale. A melspectrogram was a 2D plot of the frequency contents and their corresponding powers (represented by a color bar) of a signal as it varied with time [36]. In our experiments, the spectrogram was constructed by a bank of 34 band-pass filters as it empirically achieved the highest accuracy in testing with DL models. In Fig. 3, we illustrated melspectrogram of a representative.

3) Poincaré Plot: A Poincaré plot visualizes the 2D scatter of two consecutive data points in the sensor signal, $[x^0(l), x^0(l+1)]$ with $l \in [1, \ldots, N-1]$ [37]. A variant of Poincaré plot uses the consecutively coarse-grained time series $\bar{x}^\omega(j)$ created from the $x^0(n)$ by following formula:

$$\bar{x}^\omega(j) = \frac{1}{\omega} \sum_{n=(j-1)\omega+1}^{j\omega} x^0(n), \quad 1 \le j \le \frac{N}{\omega} \tag{3}$$

where ω is the time scale. Fig. 4 plotted a participant's kinematic data with ω is set of one.

C. Statistical Analyses

The Shapiro-Wilk test was used to verify if the data was normal distributed. In the case of normality was ensured, a one-way analysis of variance (one-way ANOVA) was implemented to determine the significance of the difference in the means of each individual characteristic between cohorts. If normality was not ensured, the Kruskal-Wallis test was conducted for the same reason. The Bonferroni test was employed to perform post-hoc analyses. At a significance threshold of $p < 0.05$, statistical results were judged to be significant. All data processing, statistical analyses, and image transforming used MATLAB (R2020b, MathWorks, USA) with an Intel Core i7, GeForce RTX 2060 SUPER (8GB VRAM) GPU, 16Gb RAM.

D. Deep Learning

1) Deep Learning Models: In this section, we provide a brief information about several representative DL models employed in the research.

a) MobileNetV2: Mobilenet is a lightweight convolutional network with 53 layers (Fig. 1C). The architecture employs depthwise separable convolutions, in which a separate convolution is applied to each of the three colour channels rather than merging and flattening them.

b) ResNet101V2: ResNet is a convolutional architecture made up of 101 layers. It implements an "identity shortcut connection" to bypasses one or more levels. This strategy aids to overcome deep networks' infamous vanishing gradient problem.

c) DenseNet: DenseNet is a variant from Densely Connected Convolution Neural Networks (CNN) with a layer will be connected feedforward to all other layers (not only adjacent layers as in the standard manner).

d) VGG16: VGG16 is a deep convolutional network alternative that has been employed in large-scale image recognition. It has 16 convolution layers, three completely linked layers, five maxpool layers, and one output softmax layer.

2) Transfer Learning: Transfer learning is a method to re-use previously obtained knowledge to resolve related ones [38]. Transfer learning is usually applied when there is too little data to train a full-scale model from scratch. It can be used to fine-tune the pre-trained models or as feature extractors. In this paper, we employ a pre-trained model as a fixed extractor, freezing the weights of the network except for the last fully connected layer. This final fully linked layer is replaced with two trainable layers, each with 100 activation nodes. Their function is to learn to tune the old features to distinguish individuals with ataxia from healthy controls. Output layer is a softmax layer with two nodes corresponding to control and ataxic cohorts.

E. Federated Learning for CA Detection

1) Federated Learning: Algorithm 1 outlines the suggesting FL scheme (FedAvg). We denote the random shuffle iteration J with holdout validation scheme of splitting 70% and 30% for training and validation sets respectively. The global training epoch index t with $t \in [1 \dots T]$. Each clinic $n \in \mathcal{N}$ participates in the training process t via the global cloud server. In each communication round t, each collaborative clinic n trains the global model with stochastic gradient descent. The actual CA raw data are stored at each local institution to ensure the data privacy. Subsequent to each local training, the performance of the global model is validated based on the local dataset and all clinics broadcast the updated θ weights to the centralised server for the model aggregation. In this study, we conduct the standard model averaging methodology FedAvg [39] to aggregate the local model weights: $\theta_{t+1} \leftarrow \frac{1}{N}\sum_1^N \theta_{n,t+1}$. From the next global epoch t, the cloud server broadcasts its new updated global weights to all member clinics to repeat the local training. The global training process continue iteratively until the average validation performance converges to a stable level.

2) Differential Privacy: The FL scheme exchanges only model parameters; nonetheless, it cannot provide sufficient protection since parts of the training data can be reconstructed via inversion of model gradients or through adversarial attacks [40]. We integrate differential privacy at each local clinic to further enhance privacy. A randomized mechanism $\mathcal{A}$: $D \rightarrow R$ is (ξ, ζ)-differentially private for $\xi > 0$ and $\zeta \in [0, 1)$ if for any two adjacent datasets $D, D' \in D$ hold that $Pr[\mathcal{A}(D) \in \mathcal{S}] \leq e^{\xi} Pr[\mathcal{A}(D') \in \mathcal{S}] + \zeta$, where $\mathcal{S} \subseteq R$. To guarantee (ξ, ζ)-differential privacy, we apply a gradient perturbation technique with differentially-private stochastic gradient descent (DP-SGD) [41]. This computes per-sample gradients, clips their ℓ_2 norm, aggregates them into a batch gradient, and adds Gaussian noise during the training. Opacus (Facebook python library, Meta AI, v1.0.0) has been employed to conduct the privacy simulations [41].

Algorithm 1 Federated Scheme Training Procedure

Global centralised server executes:
for each $j = 1, 2, \dots, J$ **do**
  Randomly shuffle and split each clinic's dataset into train (70%) and validation (30%) sets
  Initialize hyper parameters of global training epoch T, local training epoch L, local weights θ_n, learning rate σ
  for each global round $t = 1, 2, \dots, T$ **do**
    for each clinic $n \in \mathcal{N}$ **do**
      $\theta_{n,t+1} \leftarrow$ **LocalUpdate**(n, θ_t)
    end for
    $\theta_{t+1} \leftarrow \frac{1}{N}\sum_1^N \theta_{n,t+1}$
  end for
  LocalUpdate(n, θ)**:**
  for each local epoch $i = 1, 2, \dots, L$ **do**
    Compute gradient of the loss with respect to model parameters
    Perform a single optimization step
    Validate the global trained model on local validation set
    Update the local weights θ

$$\theta \leftarrow \theta - \sigma \bigtriangledown_\theta (\theta); \tag{4}$$

  end for
  Return trained θ
end for
Average validation performance over 100 iterations.

IV. Experiments and Evaluation

Experimental settings with each ML model, either in standalone, centralised or federated scheme is averaged across 150 iterations (J = *150* in Algorithm 1). Global epoch T in Algorithm 1 is 50 and model training at each local clinic uses stochastic gradient descent. Table II lists hyper parameters of the model setting. All ML experiment results were obtained by running on Pytorch (version 1.8.1) with a SuperServer SYS-1029GQ-TRT, V100-PCIE GPU (32Gb VRAM), Xeon Silver 4114 CPU @ 2.20GHz (20 core/40 threads), and 96Gb RAM.

A. Data Characteristic

First, we analysed the relevant properties of interest of each dataset using statistical tests. Fig. 5 illustrates the

TABLE II
PARAMETER SETTING

Symbol	Value
Classifier activation	Rectified Linear Unit
Classifier optimizer	Stochastic Gradient Descent
Classifier loss function	Cross Entropy Loss
Learning rate	0.001
Momentum	0.9
Numbers of global epochs	50
Numbers of local epochs	1
Batch size	16

Higher "local epochs" increases substantially the training time but not the validation accuracy.

TABLE III
VALIDATION ACCURACY OF FEDERATED DEEP LEARNING MODELS WITH RESPECT TO THREE MATRIX PATTERN TRANSFORMATIONS

	Transform technique (Accuracy %)		
Model	Recurrence	Melspectrogram	Poincaré
Resnet	81.01	64.81	73.08
Alexnet	69.10	69.25	69.00
VGG16	68.86	69.64	69.29
Squeezenet	68.91	69.26	69.20
Densenet	82.39	68.38	72.25
Shufflenet	81.24	43.83	70.27
Mobilenet	**86.69**	72.54	75.60
Mnasnet	67.64	58.59	53.65

*All models tested on image transformed from Z-axis (Anterior-Posterior) only.

characteristics of each dataset based on statistical analyses of the recurrence rate [1]. The cohorts with ataxia in each clinic differed significantly from their counterparts in the other clinics (p-$values$ highlighted in parentheses), especially in the Superior-Inferior axis with all the p-$values$ smaller than 0.05. The Medial-Lateral axis characterised with the greatest consistency between clinics with all p-$values > 0.05$ for either ataxic and control cohorts, yet the separation between ataxic and control cohorts in each clinic is significant (all p-$values < 0.05$ between two cohorts). The statistical tests may suggest employing different information from axes to obtain the best diagnostic performance which we reported separately in subsection IV-E.1.

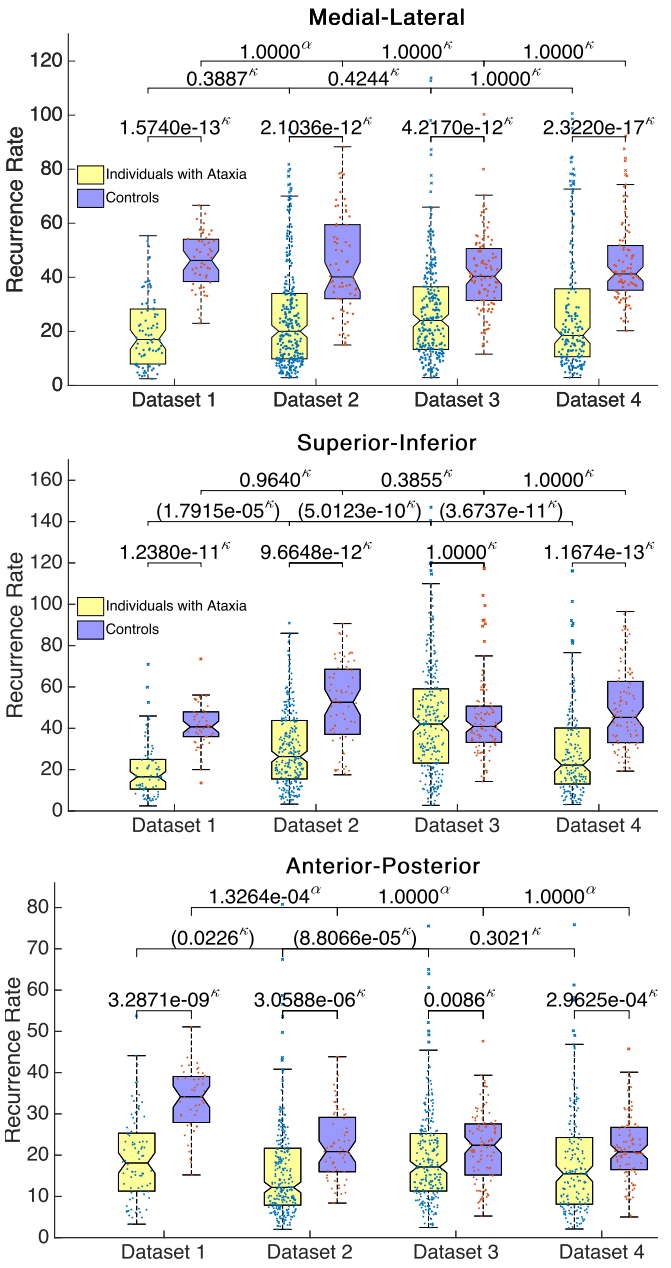


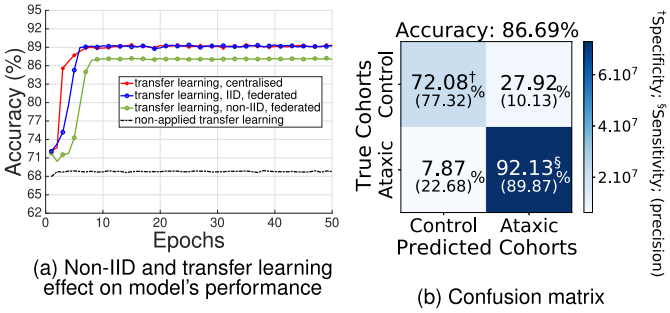
Fig. 5. Boxplots show characteristic of the dataset at each clinic based on recurrent rate and statistical analyses. κ and α denote Kruskal-Wallis and anova tests, respectively.

B. Diagnostic Results

The validation accuracy of federated DL models is compared across 50 epochs and results reported in Table III. MobileNetV2 yielded the highest diagnosis performance of 86.69% accuracy in FL and 89.30% in centralised DL, in line with the recurrence plot transformation. Compared to the centralised ML approach with 88.24% accuracy reported in [1], the DL suggested in this work yielded marginally improved (1.06%) performance. It is important to note that the testing conditions in this paper were more inferior with a diverse dataset of using two sensors, four different clinical subtests and four separate participant clinics. Among DL models, the MobileNetV2 architecture was light and simpler which resulted in faster training times. Complex and heavier networks such as the DenseNet did not produce higher accuracies.

C. Effect of Transfer Learning and IID Dataset on Performance of FL Model

In order to understand the advantage of transfer learning on small datasets, we built a simple CNN network (containing one convolution layer, max pooling, another convolution layer, and two linear transformations) and compared the performance with the proposed scheme of using transfer learning. The simulations are conducted under the centralised scheme. Fig. 6a illustrates that the employment of transfer learning is crucial for clinical application with small, limited datasets



(a) Non-IID and transfer learning effect on model's performance

(b) Confusion matrix

Fig. 6. (a) Effectiveness of transfer learning in increasing accuracy of diagnosis, illustrated in the centralised DL scheme; (b) Confusion matrix of the federated DL.

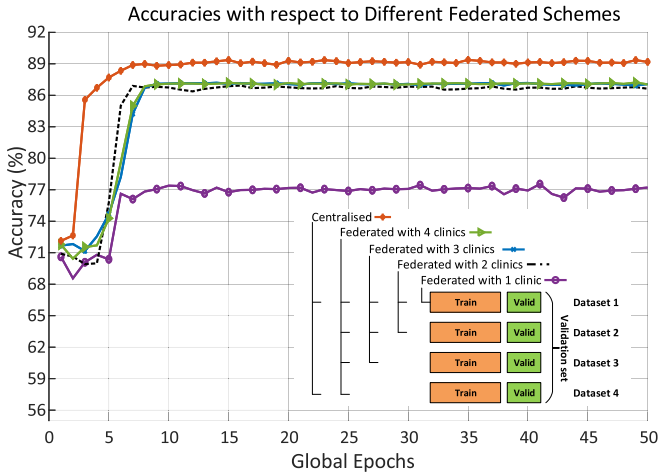


Fig. 7. Performance of federated learning in relation to the number of training clinics. The simulations used the model MobileNetV2 with recurrence plot on Z-axis.

which substantially increased the performance accuracy. It is important to note that the transfer learning performed best with some layers added on top of the DL model. In our research, we empirically used two layers with 100 hidden units per layer to produce the best performance of 86.69%.

Our CA dataset is not independent and identically distributed (non-IID), which is different in both characteristic and the number of samples. To understand the effect of the non-IID ataxia dataset on FL, we simulated the IID scenario by merging and randomly redistributing data (equally in size) to four clinics. Fig. 6a shows non-IID dataset performs inferior to the IID case (86.69% versus 89.17%, respectively) Fig. 6b presents a confusion matrix with the specificity (72.08%) and sensitivity (92.13%) of the federated DL with non-IID and transfer learning.

D. Comparing Centralised, Standalone and Federated Schemes

In order to evaluate performance of the suggested federated DL scheme, we considered the following two scenarios:

Scenario 1: We evaluated performance of FL with respect to the number of clinics participating. For a fair comparison, we created a common validation set drawn from four clinics.

TABLE IV
ACCURACY PERFORMANCE OF MOBILENETV2 WITH RESPECT TO THE NUMBER OF AXES

		Image transform techniques		
		Recurrence	Melspectrogram	Poincaré
One Axis	Me-La	80.70	69.21	71.69
	Su-In	79.48	69.20	67.90
	An-Po	**86.69**	72.97	75.98
Two Axes	Me-La & Su-In	81.87	63.67	73.14
	Su-In & An-Po	81.50	68.90	74.46
	Me-La & An-Po	83.18	69.45	74.20
Three Axes	Me-La, Su-In & An-Po	82.78	69.22	77.75

Me-la: Medial-Lateral; Su-In: Superior-Inferior; An-Po: Anterior-Posterior.

In each scheme, the clinics were selected randomly. The federated performance were plotted with increasing numbers of participating clinics as reported in Fig. 7. Centralised scheme was the traditional ML which merged all clinic datasets into one single dataset. In our dataset, despite inherent differences of ataxic cohorts across clinics (as presented by statistical analyses in Fig. 5), the performance of FL was saturated when two clinics contributed to the FL and performed comparatively good (86.69%) compared to the centralised scheme (accuracy of 89.30%). The FL model needs ten global epochs to achieve the desired performance, equivalent to 40 training epochs if the local epoch is set to one and four clinics joining the network.

Scenario 2: Here we assess the federated scheme from clinic's perspectives. The federated performance was evaluated using the global model to validate individually on each local clinic data set. Fig. 8 shows validation result of three different approaches at each clinic. The "standalone" stands for the clinic to use its own dataset, "federated" and "centralised" to use all four clinic datasets in federated and centralised scheme, respectively. This comparison answers the question whether each individual clinic gain any benefit in taking part in the federated scheme. The experimental results confirmed that joining a federated scheme would increase accuracy in 50% of clinics (2/4 institutes), 25% of the clinics exhibit no difference to the local standalone model (1/4 institutes), and 25% of the clinics decrease the accuracy (1/4 institutes). The *Dataset 2* has its performance marginally reduced due to averaging aggregation of the global model. The statistical tests in Fig. 5 also indicate CA individuals in *Dataset 2* express the largest difference with its counterparts ($p < 0.05$). Hence, joining the federated corporation would benefit the entire network, but each client should train a standalone model locally to compare and decide accordingly which model to use in practice.

E. Performance Evaluations on Federated Deep Learning

1) Evaluation of FL With Respect to the Number of Axes: We compared the accuracy of the FL with respect to different number of axes used in image transformation techniques. Table IV reported the accuracies of the FL models with the highest performance when utilizing IMU information from Anterior-Posterior with 86.69% accuracy. In this simulation,

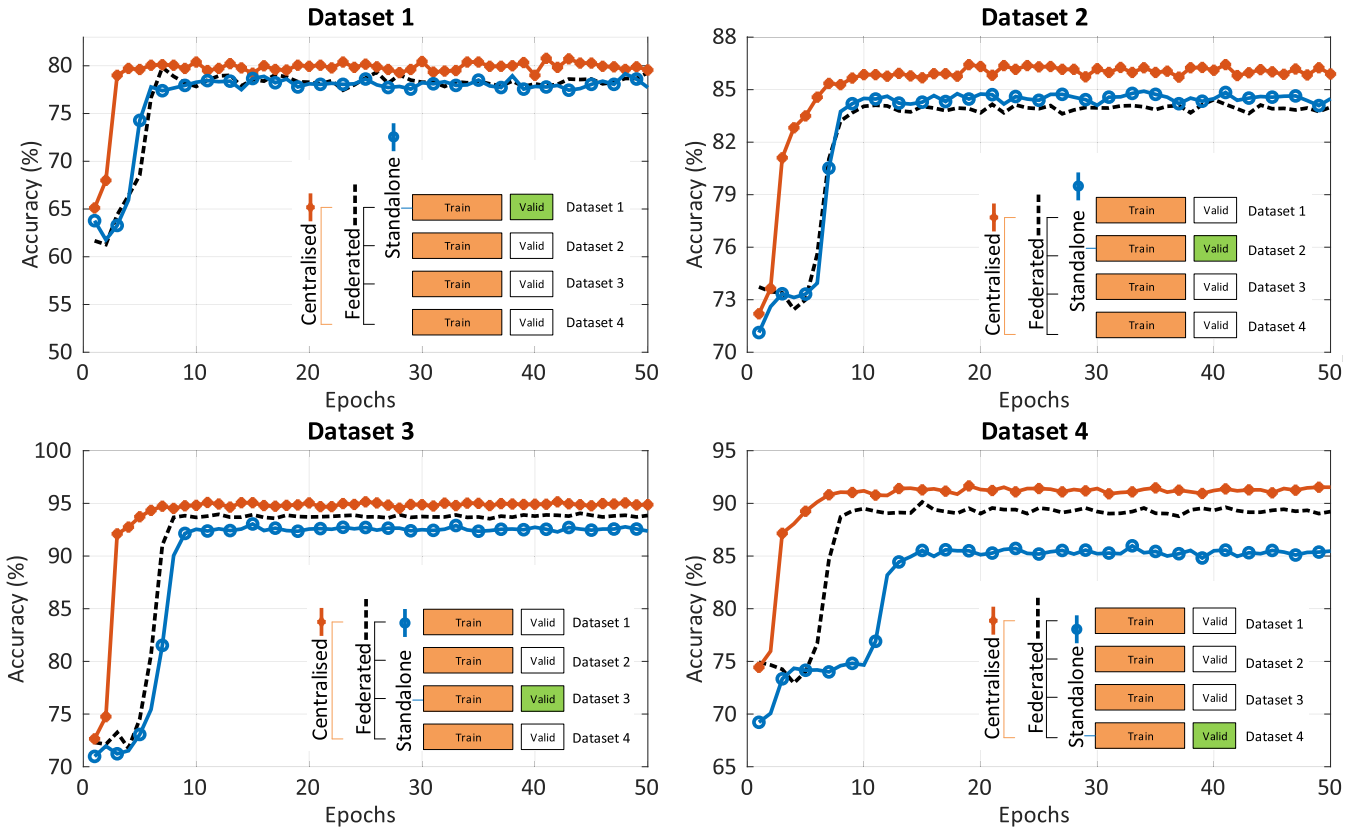


Fig. 8. Comparison centralised, standalone and federated scheme. The simulations used the model MobileNetV2 with recurrence plot on Z-axis.

TABLE V
IMPACTS OF ξ, ζ-DIFFERENTIAL PRIVACY, ON FL PERFORMANCE

Privacy parameter ξ	Max Gradient Norm 0.1	1.2	2	4	8
5	72.68	74.47	69.70	72.98	73.43
10	73.88	71.64	72.98	72.38	74.62
20	71.49	74.17	72.38	70.59	74.62
40	75.22	70.59	73.28	72.23	72.38
80	70.74	72.68	72.23	**76.71**	71.64

ζ is set to 10^{-3} as it empirically does not affect much on performance

the MobileNetV2 DL model was employed as it performed best as denoted in Table III. The recurrence plot technique superiority was consistent over Poincaré and Melspectrogram techniques across all axes. The statistical analyses in Fig. 5 shows Medial-Lateral axis with the highest separation while the Superior-Inferior axis with the lowest segregation. However, the accuracy achieved from these axes (Me-La 85.56% and Su-In 85.49%) standalone did not reflect that much difference (less than 0.07% in accuracy performance). The performance of two axes (Me-La & An-Po) and three axes were similar (88.18% and 88.65%, respectively) indicating that the model learned to ignore the less informative of Su-In axis.

2) Evaluation of Differential Privacy-Enabled FL Performance: We evaluate the FL scheme regarding the data utility performance measured by accuracy score. The privacy parameter ξ varies in range [0.01, 0.5], max_grad_norm in [0.1, 8], and ζ is set to 10^{-3}. Max_grad_norm is the maximum norm of the per-sample gradients. Any gradient with a norm higher than this will be clipped to this value. Note that smaller ξ means more privacy, more noise. Experimental results in Table V indicate the integration of an (ξ, ζ)-differential privacy solution reduces 9.98% accuracy performance.

V. CONCLUSION AND FUTURE WORKS

This study proposed a novel scheme to utilize image transformation-based approach on DL framework integrated with FL for the CA diagnosis. The results of the analysis indicated the possibility to classify 86.69% accurate using a lightweight convolutional architecture (MobileNetV2) with the recurrence plot transformer. Compared to traditional ML approach, applying DL achieved a better accuracy with higher flexibility to engage with different datasets. It also saved analysis and deployment time by eliminating the laborious work of feature extraction. The FL enables privacy protection for the participant clinics with a deployable and practical scheme for implementation. Future work will involve stacking two or more distinct models with other pattern matrix transformation techniques to enhance overall performance. We will also secure the FL using blockchain and utilise tailored models for texture datasets.

ACKNOWLEDGMENT

The authors thank Sarah Milne (Murdoch Children's Research Institute, Melbourne), Jillian Chua (Ryde Hospital,

Sydney), Amy Robinson (Ryde Hospital, Sydney), Shannon Williams (Royal Hospital, Perth), Kristen Grove (Sir Charles Gairdner Hospital, Perth), and Hannah Ross (Monash Health, Melbourne) for their support in the training all the physiotherapists, recruiting participants, collecting, and organising data from all the sites.

REFERENCES

[1] T. Ngo *et al.*, "Balance deficits due to cerebellar ataxia: A machine learning and cloud-based approach," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1507–1517, May 2021.

[2] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[3] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinf.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[4] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, Dec. 2022.

[5] M. Hao, H. Li, X. Luo, G. Xu, H. Yang, and S. Liu, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6532–6542, Oct. 2020.

[6] Y. Liu, J. J. Q. Yu, J. Kang, D. Niyato, and S. Zhang, "Privacy-preserving traffic flow prediction: A federated learning approach," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7751–7763, Aug. 2020.

[7] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent IoT applications: A cloud-edge based framework," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 35–44, 2020.

[8] Y. Zhao, H. Liu, H. Li, P. Barnaghi, and H. Haddadi, "Semi-supervised federated learning for activity recognition," 2020, *arXiv:2011.00851*.

[9] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for Internet of Things: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1622–1658, 3rd Quart., 2021.

[10] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, early access, Dec. 16, 2020, doi: 10.1109/TMC.2020.3045266.

[11] N. Rieke *et al.*, "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–7, Dec. 2020.

[12] R. Kumar *et al.*, "Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging," *IEEE Sensors J.*, vol. 21, no. 14, pp. 16301–16314, Jul. 2021.

[13] S. Warnat-Herresthal *et al.*, "Swarm learning for decentralized and confidential clinical machine learning," *Nature*, vol. 594, no. 7862, pp. 265–270, 2021.

[14] Z. Yang, S. Zhong, A. Carass, S. H. Ying, and J. L. Prince, "Deep learning for cerebellar ataxia classification and functional score regression," in *Machine Learning in Medical Imaging*, G. Wu, D. Zhang, and L. Zhou, Eds. Cham, Switzerland: Springer, 2014, pp. 68–76.

[15] Z. Chang *et al.*, "Accurate detection of cerebellar smooth pursuit eye movement abnormalities via mobile phone video and machine learning," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Dec. 2020.

[16] B. Kashyap, P. N. Pathirana, M. Horne, L. Power, and D. Szmulewicz, "Quantitative assessment of speech in cerebellar ataxia using magnitude and phase based cepstrum," *Ann. Biomed. Eng.*, vol. 48, no. 4, pp. 1322–1336, Apr. 2020.

[17] H. Tran, P. N. Pathirana, M. Horne, L. Power, and D. J. Szmulewicz, "Quantitative evaluation of cerebellar ataxia through automated assessment of upper limb movements," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 1081–1091, May 2019.

[18] J. Lee, Y. Kagamihara, and S. Kakei, "A new method for functional evaluation of motor commands in patients with cerebellar ataxia," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0132983.

[19] C. Hohenfeld *et al.*, "Application of quantitative motor assessments in friedreich ataxia and evaluation of their relation to clinical measures," *Cerebellum*, vol. 18, no. 5, pp. 896–909, Oct. 2019.

[20] K. Bando, T. Honda, K. Ishikawa, Y. Takahashi, H. Mizusawa, and T. Hanakawa, "Impaired adaptive motor learning is correlated with cerebellar hemispheric gray matter atrophy in spinocerebellar ataxia patients: A voxel-based morphometry study," *Frontiers Neurol.*, vol. 10, p. 1183, Nov. 2019.

[21] K. D. Nguyen, P. N. Pathirana, M. Horne, L. Power, and D. J. Szmulewicz, "Entropy-based analysis of rhythmic tapping for the quantitative assessment of cerebellar ataxia," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101916.

[22] M. Ghislieri, L. Gastaldi, S. Pastorelli, S. Tadano, and V. Agostini, "Wearable inertial sensors to assess standing balance: A systematic review," *Sensors*, vol. 19, no. 19, p. 4075, Sep. 2019.

[23] T. Honda *et al.*, "Assessment and rating of motor cerebellar ataxias with the Kinect v2 depth sensor: Extending our appraisal," *Frontiers Neurol.*, vol. 11, p. 179, Mar. 2020.

[24] A. Prochazka *et al.*, "Deep learning for accelerometric data assessment and ataxic gait monitoring," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 360–367, 2021.

[25] D. Phan, N. Nguyen, P. N. Pathirana, M. Horne, L. Power, and D. Szmulewicz, "A random forest approach for quantifying gait ataxia with truncal and peripheral measurements using multiple wearable sensors," *IEEE Sensors J.*, vol. 20, no. 2, pp. 723–734, Jan. 2020.

[26] B. Müller, W. Ilg, M. A. Giese, and N. Ludolph, "Validation of enhanced Kinect sensor based motion capturing for gait assessment," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0175813.

[27] D. Jarchi, J. Pope, T. K. M. Lee, L. Tamjidi, A. Mirzaei, and S. Sanei, "A review on accelerometry-based gait analysis and emerging clinical applications," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 177–194, 2018.

[28] B. Roche *et al.*, "Test-retest reliability of an instrumented electronic walkway system (GAITRite) for the measurement of spatio-temporal gait parameters in young patients with Friedreich's ataxia," *Gait Posture*, vol. 66, pp. 45–50, Oct. 2018.

[29] U. M. Küng *et al.*, "Postural instability in cerebellar ataxia: Correlations of knee, arm and trunk movements to center of mass velocity," *Neuroscience*, vol. 159, no. 1, pp. 390–404, Mar. 2009.

[30] X. Yang *et al.*, "*S*-band sensing-based motion assessment framework for cerebellar dysfunction patients," *IEEE Sensors J.*, vol. 19, no. 19, pp. 8460–8467, Oct. 2018.

[31] K. D. Nguyen, L. A. Corben, P. N. Pathirana, M. K. Horne, M. B. Delatycki, and D. J. Szmulewicz, "The assessment of upper limb functionality in friedreich ataxia via self-feeding activity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 924–933, Apr. 2020.

[32] R. Krishna, P. N. Pathirana, M. K. Horne, L. A. Corben, and D. J. Szmulewicz, "Quantitative assessment of friedreich ataxia via self-drinking activity," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1985–1996, Jun. 2021.

[33] K. Z. Gajos *et al.*, "Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection," *Movement Disorders*, vol. 35, no. 2, pp. 354–358, Feb. 2020.

[34] E. H. Buder, "Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990," *Voice Qual. Meas.*, vol. 119, p. 244, Jan. 2000.

[35] C. L. Webber and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," *J. Appl. Physiol.*, vol. 76, no. 2, pp. 965–973, Feb. 1994.

[36] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2010.

[37] T. S. Henriques, S. Mariani, A. Burykin, F. Rodrigues, T. F. Silva, and A. L. Goldberger, "Multiscale Poincaré plots for visualizing the structure of heartbeat time series," *BMC Med. Informat. Decis. Making*, vol. 16, no. 1, pp. 1–7, Dec. 2015.

[38] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[39] S. Banerjee, R. Misra, M. Prasad, E. Elmroth, and M. H. Bhuyan, "Multi-diseases classification from chest-X-ray: A federated deep learning approach," in *AI 2020: Advances in Artificial Intelligence*, M. Gallagher, N. Moustafa, and E. Lakshika, Eds. Cham, Switzerland: Springer, 2020, pp. 3–15.

[40] M. Y. Lu *et al.*, "Federated learning for computational pathology on gigapixel whole slide images," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102298.

[41] A. Yousefpour *et al.*, "Opacus: User-friendly differential privacy library in PyTorch," 2021, *arXiv:2109.12298*.