Computer Methods and Programs in Biomedicine 242 (2023) 107840

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

A deep learning-based telemonitoring application to automatically assess oral diadochokinesis in patients with bulbar amyotrophic lateral sclerosis

Lucia Migliorelli [a,b,*], Lorenzo Scoppolini Massini [a,b], Michela Coccia [c], Laura Villani [d], Emanuele Frontoni [e,b,f], Stefano Squartini [a]

[a] Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy
[b] AIDAPT S.r.l., Ancona, Italy
[c] Centro Clinico NeuroMuscular Omnicentre (NeMO), Fondazione Serena Onlus, Ancona, Italy
[d] Department of Neuroscience, Neurorehabilitation Clinic, Azienda Ospedaliero-Universitaria delle Marche, Ancona, Italy
[e] Department of Political Science, Communication and International Relations, Università degli Studi di Macerata, Macerata, Italy
[f] Nemo Lab, Milan, Italy

ARTICLE INFO

Keywords:
Oral diadochokinesis
Deep learning
Bulbar-onset amyotrophic lateral sclerosis
Dysarthria

ABSTRACT

Background and objectives: Timely identification of dysarthria progression in patients with bulbar-onset amyotrophic lateral sclerosis (ALS) is relevant to have a comprehensive assessment of the disease evolution. To this goal literature recognized the utmost importance of the assessment of the number of syllables uttered by a subject during the oral diadochokinesis (DDK) test.

Methods: To support clinicians, this work proposes a remote deep learning-based system, which consists (i) of a web application to acquire audio tracks of bulbar-onset ALS patients and healthy control subjects while performing the oral DDK test (*i.e.,* repeating the /pa/, /pa-ta-ka/ and /oo-ee/ syllables) and (ii) a DDK-AID network designed to process the acquired audio signals which have different duration and to output the number of per-task syllables repeated by the subject.

Results: The DDK-AID network overcomes the comparative method achieving a mean Accuracy of 90.23 in counting syllables repeated by the eleven bulbar-onset ALS-patients while performing the oral DDK test.

Conclusions: The proposed remote monitoring system, in the light of the achieved performance, represents an important step towards the implementation of self-service telemedicine systems which may ensure customised care plans.

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease of adulthood (the average age of sufferers is 58-60 years old). It is the most common motor neuron disease and has a prevalence of about 0.6 and 3.8 per 100000 person-years [1,2]. ALS causes the gradual loss of spinal, bulbar and cortical motor neurons, leading to paralysis of voluntary muscles and even respiratory ones [3].

The occurrence of ALS-related bulbar symptoms, such as dysarthria, denotes a bulbar involvement in the ALS evolution and it represents a crucial milestone in the development of the bulbar-onset ALS. Dysarthria encompasses a range of neurological speech disorders characterized by irregularities in the strength, speed, range, steadiness, tone, or precision of movements involved in breathing, phonation, resonance, articulation, and prosody [4]. Thus, this sign has devastating consequences on an individual's ability to communicate and quality of life [5].

Identifying the evolution of dysarthria has a pivotal clinical value: to predict the progress of the disease, to prescribe compensatory strategies (*e.g.,* assistive communication devices) for ensuring those who are affected to live as well as possible, and to find new outcome measures for clinical trials [6]. Although dysarthria assessment holds significant importance, it primarily relies on visual observation by clinicians and the usage of the Robertson dysarthria profile. This is a clinical rating scale with different items to assess various aspects – such as breathing, phonation, facial musculature, diadochokinesis (DDK), reflexes, articulation, intelligibility and prosody – as to determine the extent of impairment and plan an appropriate treatment plan [7]. Among the tasks from the

[*] Corresponding author at: Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy.
E-mail address: l.migliorelli@staff.univpm.it (L. Migliorelli).

https://doi.org/10.1016/j.cmpb.2023.107840
Received 22 August 2022; Received in revised form 15 September 2023; Accepted 1 October 2023

Available online 5 October 2023

Robertson Dysarthria Profile, the longitudinal evaluation of speaking rate during the test of oral DDK is of utmost importance to map the dysarthria progression [8–10]. DDK is a physically-demanding test dealing with the rapid repetition of syllables in a time interval ranging from 30 seconds to one minute [11]. The procedure requires both the repetition of single syllables (as: /pa/), also known as alternating motion rate, or polysyllables (as: /pa-ta-ka/, or /oo-ee/), also referred as sequential motion rate [12,13]. Uttering these syllables (/pa/, /ta/, /ka/, /oo-ee/) require bilabial, dental, velar and vowel actions. As a consequence, possible difficulties in conducting DDK test over time, such as a reduced speaking rate with respect to the previous time, may timely reveal the presence of bulbar ALS-related oral-motor deficits [8,6,14].

Despite its clinical relevance, oral-DDK test is mainly performed during the outpatient assessment and the clinicians directly evaluate patients by counting the number of syllables repeated by the subject for each of the tasks [15]. This procedure is sporadic and suffers from fatigability bias, induced in the patients by the travel to the facility which highly influences their performance during the evaluation [5,16,17]. Moreover, the assessments are often collected in paper format, jeopardising data availability, sharing and longitudinal consulting [6,18]. Computer-assisted methodologies, based on the analysis of audio recordings, have been proposed in literature to automatically evaluate patients' performance in the oral DDK test [19,20]. However, these systems are designed to operate in a controlled environment and with high-performance audio acquisition systems.

With the view to support clinicians in the automatic and close monitoring of dysarthria, the proposed work presents a convolutional neural network (CNN)-based system to process audio signals and remotely evaluate the oral-DDK test. The system is validated on recordings from patients with bulbar-onset ALS and healthy control subjects while performing the three tasks of repeating the /pa/, /pa-ta-ka/ and /oo-ee/ syllables for 30 seconds. The subjects involved in the study performed the test at home with consumer acquisition devices – *i.e.,* earphones connected to smartphones, tablets or PCs – and the CNN automatically counted the number of syllables repeated for each of the tasks. The innovative contributions of the work are the following:

- The implementation of a custom network architecture, namely DDK-AID network, inspired by the general-purpose object detector in [21]. The DDK-AID network, unlike the one proposed in [21], (i) reduces the size of the bounding box detection from 2D (width and height) to 1D (width), to save computation and (ii) inputs audio signals variable in duration.
- The implementation of an on-the-fly methodology to create synthetic training-audio signals, from real ones, of people performing the oral DDK test. The aim of this methodology was to tackle the variability in both bulbar-onset ALS population's vocal performance and the use of consumer devices for acquisitions, all while enhancing the network's generalization capabilities when applied to recordings obtained in uncontrolled scenarios [19].
- The proposal of a remotely self-usable prototype system to perform the oral DDK test at home, consisting of a web-application that the subjects involved in the study can use on their personal device.

Pursuing studies on self-service telemedicine systems has, of all, the value of: broadening the possibilities of research, favouring the collection of innovative assessment indexes for clinical trials, and actively involving patients in their care plan, allowing them to carry out assessments in a familiar environment while feeling almost in touch with their clinical reference.

The rest of the paper is organized as follows: Sec. 2 presents the relevant state-of-the-art contributions in the field of oral DDK-test assessment in dysarthric subjects. Sec. 3 presents the implemented methodology while technical details for enabling fair comparisons are shown in Sec. 4. The results achieved are detailed in Sec. 5 and discussed in Sec. 6. Sec. 7 concludes the work and proposes possible future developments.

2. State of the art

In the past decades, some computer-based approaches were developed to support clinicians in assessing patients' performance while executing the oral DDK test.

Rong in [6] proposes a method to automatically monitor ALS patients while performing the oral-DDK test. The approach is based on the extraction of vocal features from the audio signal and the subsequent application of wavelets to map the temporal pattern of syllable repetitions.

Similarly in [22,23], the authors apply filtering techniques before implementing both a selective-search algorithm to identify signal peaks and clustering to discern those most likely to be a DDK-repetition.

These approaches were validated for ALS-dysarthric patients. However both acquisition protocols require audio signals to be recorded in quiet rooms, with performant microphones (*e.g.,* head-mounted condenser microphone) and in controlled scenarios (*i.e.,* with supervision by clinicians). Indeed, the implemented audio-processing methodologies mainly rely on standard signal-processing techniques which may be unsuitable for tackling highly variable audios in a dataset *i.e.,* acquired with consumer devices and in uncontrolled scenarios typical of telemedicine systems [24].

Inspired by recent considerations that showed the potentiality of deep learning over standard signal-processing techniques when dealing with multimedia data in closer fields of research [25,26], in [19] the authors implement a semi-automatic system consisting of two subsequent CNNs. The experimental approach, designed for dysarthric patients but tested on healthy control subjects only, assesses subjects' performance while carrying the DDK-test out via a sliding-windowing algorithm. Then, a corrective module enables the manual refinement of networks' predictions. Despite the breakthrough in implementing deep-learning procedures, the semi-automatic nature of the approach always requires the intervention of a clinician. Furthermore, the use of two subsequent networks has been outperformed by the implementation of individual multi-task frameworks [27].

With the view to overcome limitations posed in [19], in [28], the authors implement a fully-automatic framework based on Faster R-CNN to assess DKK-test in patients suffering from multiple sclerosis. The proposed approach is computationally burdensome (number of trainable parameters ∼60 M) and, when tested on bulbar-onset ALS patients, performance was not satisfactory. Indeed, as proven in [23], the complexity of the network may bring the training data to be memorized. This causes the network to lose its effective capacity in handling the high variability inherent in data in terms of vocal performance and vocal alterations induced by pathology trajectory [29].

It is worth noting that these two latter approaches, as the previous ones [6,22,23], handle signals acquired in a controlled scenario, moreover, they process audio signals having the same duration as to have inputs equal in size for the fully connected (FC) layers. These strict requirements may hamper the translation of such computer-assisted technologies in the actual clinical practice while hardening the possibility of integrating the methods in a telemedicine application.

To overcome the state-of-the-art limitations, this paper presents a system based on deep learning to remotely assess the number of syllables repeated by patients with bulbar-onset ALS and healthy control subjects while performing the oral DDK test. Inspired by work in closer fields [28] our system exploits the potential of deep learning to cope both with the complexities of data acquired with consumer microphones and the audio-data variability induced by dysarthria staging in the involved patients.

Unlike other state-of-the-art contributions, our DDK-AID model has a lower computational complexity – and consequently lower costs – with the view to deploy such a monitoring application within a cloud

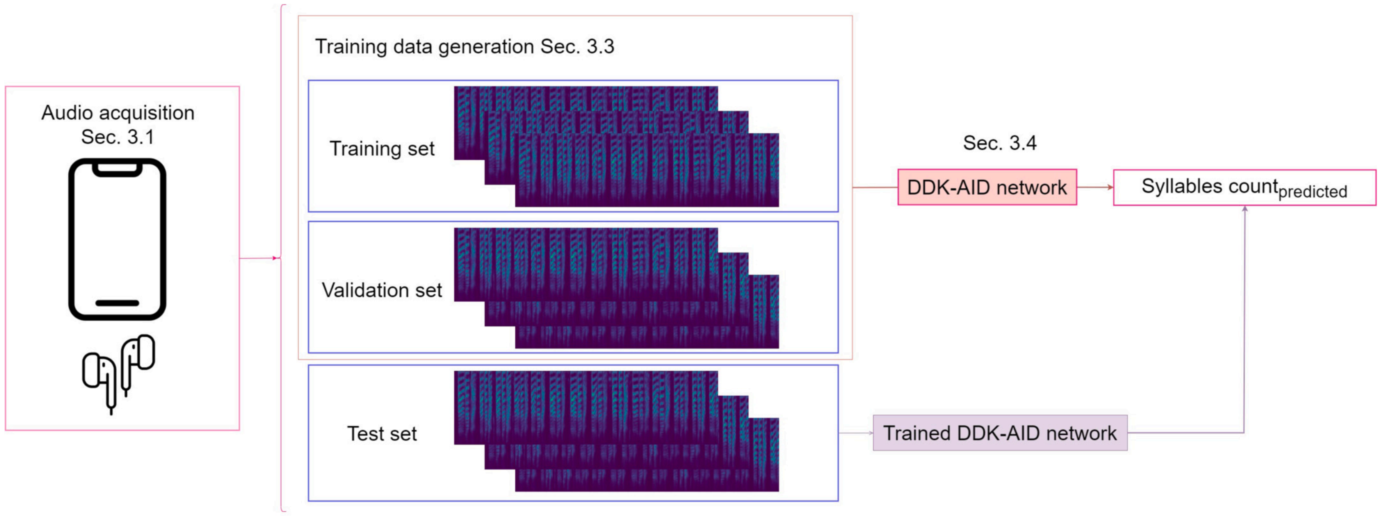


Fig. 1. Workflow of the convolutional neural network (CNN)-based system to remotely assess the performance of healthy control and bulbar-onset amyotrophic lateral sclerosis (ALS) subjects who carry out the oral diadochokinesis (DDK)-test.

infrastructure [30]. Moreover, the network was trained and validated from audio recordings (i) of different duration, following guidelines in [31] and (ii) acquired with commonly used devices and in uncontrolled environments (each subject was free to use the application at home without clinician's supervision). The workflow of the proposed system is shown in Fig. 1.

3. Methods

3.1. Data acquisition

Audio recordings were acquired from 32 healthy control subjects and 11 bulbar-onset ALS patients while carrying out the oral-DDK test.

The acquisitions were made via the Homely Care web application (Fig. 2) that the subjects involved in the study can use on their personal device such as smartphone, computer, tablet. Homely Care enables subjects to sequentially perform the 3 oral-DDK tasks of repeating: (i) the /pa/ syllable (ii) the /oo-ee/ syllables and (iii) the /pa-ta-ka/ syllables, while recording the audios. Following the protocol in [31], each task can last a maximum of 30 seconds and the subject, eventually feeling fatigued, has the possibility to stop the acquisition at any time.

As shown in Fig. 2, the first page of the application explains to the user how to perform the test optimally. For example, the user is asked to wear headphones with a microphone to capture qualitatively better audio recordings. Tutorials with a speech language pathologist while performing the oral-DDK test are provided too as well as instructions to start a recording.

The application also allows the user to eventually discard and repeat the audio acquisition if adverse events occurred such as the sudden ringing of the telephone or intercom.

During the execution of the test, the user was reminded to: (i) acquire in a room without TV and radio turned on, (ii) wear earphones with microphone.

Then, the acquired audio recordings were safely stored on the cloud architecture and analysed via the DDK-AID network.

As showed in Table 1, a total of 350 audio acquisition (approximately 58 minutes of recordings) were initially acquired via the Homely Care web application. These acquisitions are the result of a first-step of dataset cleaning: some recordings (*e.g.,* 5 recordings of /pa-ta-ka/ syllables-repetition task) were excluded due to excessive background noise, in most cases caused by a person talking loudly in background.

The total number of acquisitions made by the subjects involved in the study is shown in Fig. 3 for the /pa-ta-ka/ syllables-repetition task. Similar trends occur for the other two oral-DDK tasks. As visible from the pie chart, healthy control subjects perform the DDK task only once, whereas bulbar-onset ASL patients were left to use the Homely Care application free for 4 months.

Table 1
Total number of recordings acquired through the Homely Care application for each of the 3 oral diadochokinesis (DDK) tasks.

	/pa/	/pa-ta-ka/	/oo-ee/
Bulbar-onset ALS patients	97	92	97
Healthy control subjects	17	32	15

3.2. Data preprocessing

Inspired by works in closer fields [28,32,33], these audio acquisitions are processed by the DDK-AID network as mel spectrograms (S) extracted from the audio recorded at a sample rate of 8000 Hz. The S is calculated over 64 bands using a Hann window of 0.064 s length and a 25% shift. To accentuate the energy of the syllables in frequency we applied the following equation obtaining an enhanced S (S_e):

$$S_e = \sqrt[3]{10\log(S(\tau,\nu)+1)}\ \forall\tau\in[0,T],\forall\nu\in[0,64] \quad (1)$$

Where $S(\tau,\nu)$ is a scalar entry of the S at time (τ) and frequency band (ν). T is the total duration of the acquisition.

The result obtained was then normalised as follows:

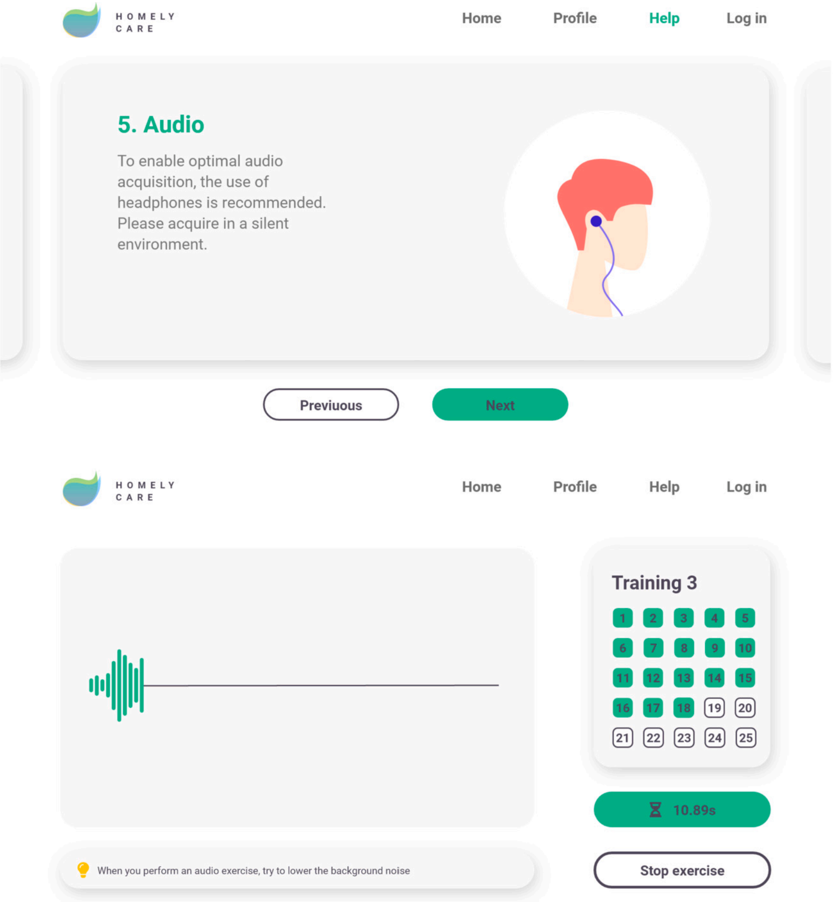
$$S_N = \frac{S_e(\tau,\nu)-median(S_e)}{percentile(S_e,95)}\ \forall\tau\in[0,T],\forall\nu\in[0,64] \quad (2)$$

3.3. Training data generation

The choice to let patients use the application freely led to the collection of an unbalanced number of acquisitions per subject (see Table 1 and Fig. 3).

Therefore, to lower the risk of overfitting [34] during training, for each subject involved, we decided to randomly extract one single audio recording per task. This has resulted in a pre-training dataset of approximately 21 minutes. The original 58-minutes dataset partitioning is outlined in Fig. 4.

To make up for the lack of data and to increase data variability, these 21 minutes of recordings were used as a prior to implement a methodology to generate the actual training dataset. This dataset consists entirely of generated audio signals both for bulbar-onset ALS



5. Audio

To enable optimal audio acquisition, the use of headphones is recommended. Please acquire in a silent environment.

Training 3

When you perform an audio exercise, try to lower the background noise

Fig. 2. Example screenshot of the Homely Care web application, enabling users to perform the oral diadochokinesis (DDK) test. The screen on top shows the starting page of the application. This has all the information the users need to perform the oral DDK test at home, *e.g.,* it invites the users to wear headphones to ameliorate the quality of the audio acquisition, offers the users an audio tutorial on how best to carry out the test. On the bottom the actual acquisition screen is shown. The users have the option to press play when are ready to perform the test. They can stop when they want (considering that the test lasts a maximum of 30 seconds). When performing the oral DDK test, the users are always advised to record in non-noisy environments. In addition, the users have the option of deleting the recording and repeating the test if a sudden noise or adverse event has occurred.

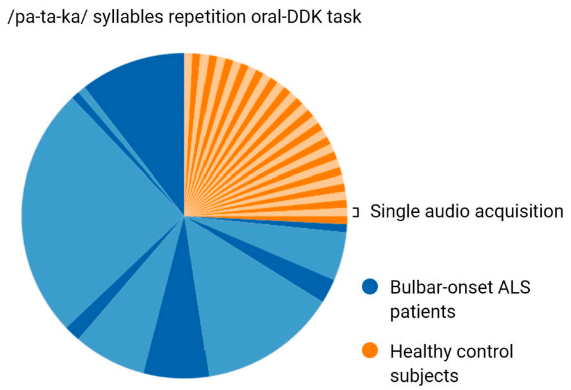


Fig. 3. Pie chart showing the number of acquisitions for the /pa-ta-ka/ repetition task produced by the subjects involved in the study. The larger the slice area, the higher the number of acquisitions conducted by the subject. Blue denotes bulbar-onset amyotrophic lateral sclerosis (ALS) patients (the two distinct shades of blue identify individual patient's acquisitions), orange denotes healthy control subjects (the two distinct shades of orange identify individual control subject's acquisition). As visible, healthy control subjects performed a single oral DDK-test recording. In contrast, bulbar-onset ALS patients were free to use the application, and the areas of the blue pie chart slices are consequently variable.

patients and healthy control subjects (Fig. 4). The training-dataset generation follows the procedure described below:

- As shown in Fig. 5, from each spectrogram of subject i two groups of portions were selected. The one with the syllable repetition $F_i = \{f_{i,0}, f_{i,1}, \dots, f_{i,d}\}$ and that with pause between two consecutive repetitions $E_i = \{e_{i,0}, e_{i,1}, \dots, e_{i,m}\}$. Where d and m are the total number of syllables and pauses noted in a single audio recording, respectively.
- To maximise the variability of the generated spectrograms – and consequently mitigate overfitting – all the E_i were merged in a single group $E = \{e_0, e_1, \dots, e_z\}$, without distinguishing the subject from whom they were recorded. In contrast, $f_{i,j}$ portions were kept separate by subject. z represents the total number of pauses annotated in the pre-training dataset.
- During the training, spectrograms were on-the-fly generated by alternately concatenating the previously cut $f_{i,j}$ and e_j portions (Fig. 6). This procedure enables, for each subject, to generate audio signals lasting differently and with intervals between consecutive syllables repetition variable in duration as well, simulating possible oral DDK test-induced fatigue. The per-task spectrograms generation algorithm runs on-the-fly at every batch as to allow the network to handle new data each time. Its flow is illustrated below:
  1. **Choice of subject which F_i portions belong to.** Considering that, the original audio signals have variable number of syllable repetitions per time (*e.g.,* minimum 8 to maximum 60 for the /pa-ta-ka/ syllables repetition), a weighted random strategy is adopted to favour subjects with a higher number of syllable

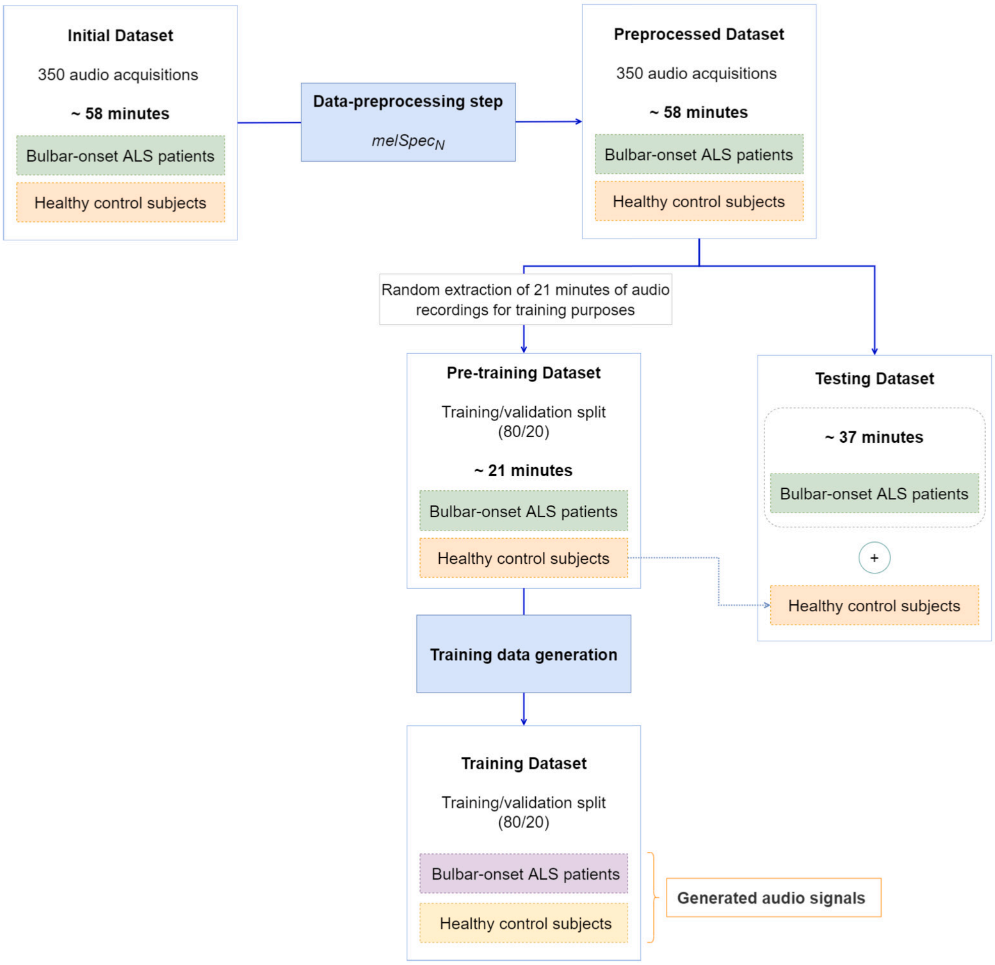


Fig. 4. Original dataset partitioning. Considering the increased number of audio acquisitions made by bulbar-onset ALS patients, from the original 58-minutes pre-processed dataset we randomly derived two sets: (i) a first set -namely pre-training set- with 21 minutes of audio recordings from bulbar-onset amyotrophic lateral sclerosis (ALS) patients and healthy control subjects and (ii) a second one with the remaining 37 minutes of recordings from bulbar-onset ALS patients. The 21 minutes of audio recordings serve as a prior for generating fictitious audio signals from training and validation purposes. While we test the performance of the network on the 37 minutes of recordings from bulbar-onset ALS patients and on the original recordings of healthy control subjects in the pre-training set. We would like to emphasise that no original signals were used in the training and validation phase, but were only generated with the procedure *Training data generation*.

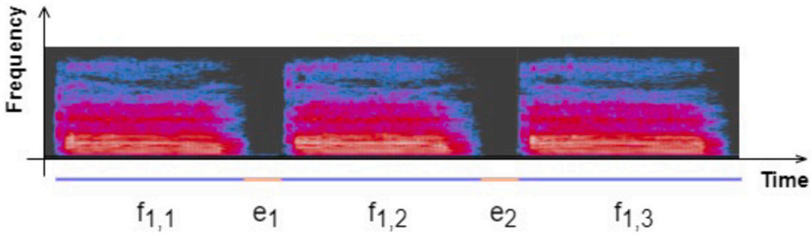


Fig. 5. Spectrogram partition consisting of three /pa/ syllables repetition. The $f_{i,j}$ portions delimite the duration of the /pa/ syllable while the e_j portions identify the time between the repetition of two successive /pa/ syllables. i is an index that identifies a specific subject while j is a progressive index to distinguish different /pa/ syllables within the audio.

repetitions. In particular, at each iteration of the algorithm, the i-th subject had a probability p_i of being selected equal to:

$$p_i = \frac{|F_i|}{\sum_{k=0}^{n} |F_k|} \tag{3}$$

Where $|F_i|$ represents the cardinality of the set (*i.e.,* the number of repetitions of the i subject) and n the number of subjects involved in the study.

2. **Choice of number of f-repetitions to be included in each generated signal.** The number of repetitions (namely s) is randomly fixed each time to generate the signal.
3. **Definition of the average duration of the f-portions.** Starting from the set F_i, a single portion $f_{i,j}$ is randomly extracted. The duration of this portion ($Len(f_{i,j})$) constrains the average duration of the subsequent generated portions. In particular each portion of the generated spectrogram is stretched based on a Gaussian-distribution procedure as follows:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4}$$

Where σ is experimentally set to 0.1 and μ is equal to $Len(f_{i,j})$

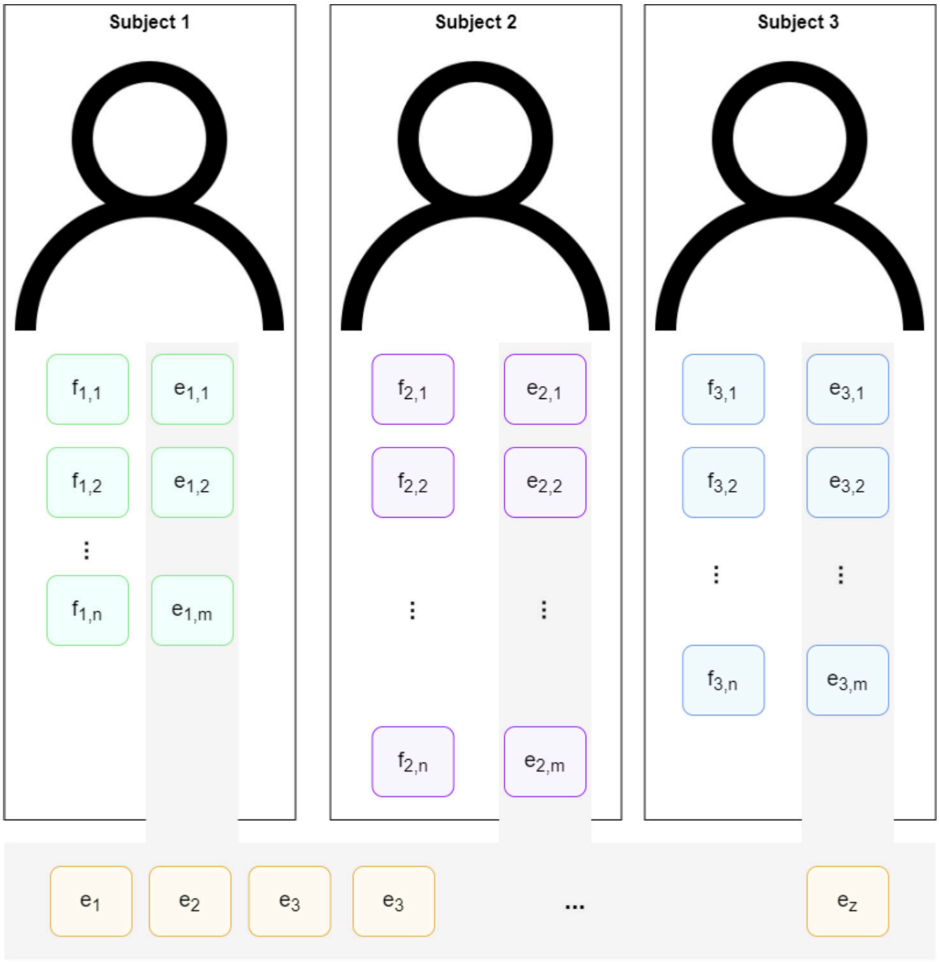


Fig. 6. On-the-fly training audio-signals generation. $f_{i,j}$ portions, the portions of the signal with the syllable repetition, were kept separate by subject while e_j ones, the portions of the signal between two subsequent repetitions, were kept together.

4. **Alternating choice of e_j- and $f_{i,j}$-portions.**
   1 A portion e_j is randomly selected from E.
   2 The chosen e_j is stretched to $\tilde{e}_j$ with a duration $Len(\tilde{e}_j)$ equal to:

$$Len(\tilde{e}_j) = \alpha G(x)\,|_{x\in\mathbb{R}} \tag{5}$$

   Where α is a number between 0 and 1 which was set following experimental assessment on real signals. It serves to make the silence instants of the generated signal consistent with those of the real one.
   2a To simulate subject's breaks, which are particularly frequent in bulbar-onset ALS patients, as experimentally verified by the analysis of the collected dataset, the 5% of the cases followed the equation below:

$$Len(\tilde{e}_j) = \beta G(x)\,|_{x\in\mathbb{R}} \tag{6}$$

   Where β was experimentally set equal to 10.
   3 A portion $f_{i,j}$ is randomly selected from F_i.
   4 The chosen $f_{i,j}$ is stretched to $\tilde{f}_{i,j}$ making its duration equal to $Len(\tilde{f}_{i,j})$, according to the following equation:

$$Len(\tilde{f}_{i,j}) = G(x)\,|_{x\in\mathbb{R}} \tag{7}$$

In Fig. 7 are shown both generated (top) and real (bottom) audio spectrograms from two different subjects.

3.4. DDK-AID network

A model inspired by You Only Look Once (YOLO) X [21] network was used to tackle the DDK-test assessment. The network, showed in Fig. 8, has 7 convolutional layers followed by 2 FC layers. As in [21], the DDK-AID Network is an anchor-free CNN. Excluding the anchoring mechanism was driven by two motivations: (i) to avoid a clustering analysis to determine a set of optimal anchors and (ii) to reduce the complexity of the detection heads and the number of per-image predictions. Unlike YOLO X, the DDK-AID network is originally designed to handle inputs of variable size along the temporal dimension. This is relevant to the task we deal with, where subjects are left free to perform the DDK-test for a maximum of 30 seconds possibly stopping when they get tired.

Indeed the DDK-AID network takes in input the original spectrogram with dimension $(w,\ h,\ 1)$ where w are the signal's frames and varies from acquisition to acquisition and h are the bands (*i.e.*, 64). Prior to entering the first convolutional layer, the spectrogram undergoes a padding operation aimed at making the input dimension w a multiple of 128. This way, the new padded spectrogram dimension becomes $(w+p,\ h,\ 1)$ with $w+p = 128k, k \in \mathbb{N}$. Controlling the padding operation at the input level -and not at each convolutional layer- allows for a 1-to-1 correspondence with the output produced by the network, enabling DDK-AID predictions to be consistent in size with respect to the original-spectrogram annotations, simplifying loss management.

Then, the padded spectrogram enters 7 subsequent convolutional layers. Each of these layers is activated by a Leaky Rectified Linear Unit (ReLU) and implements batch normalization. The convolutional layers

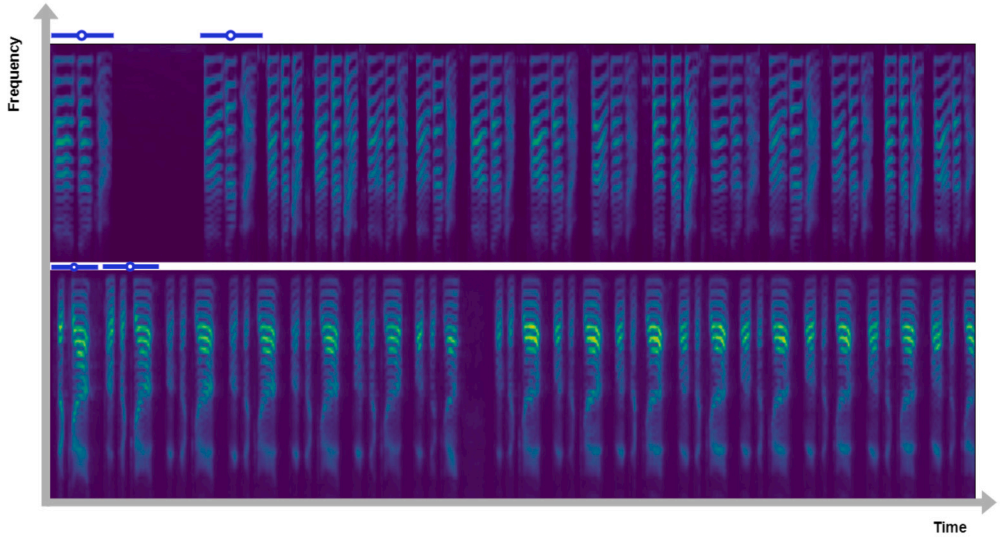


Fig. 7. The spectrogram shows the generated /pa-ta-ka/ syllables repetition (at the top of the figure) and real one (at the bottom of the figure). Sample of annotations (*i.e.,* midpoint and length of the segment delimiting the syllables repetition) for training and validation sets are shown in blue.

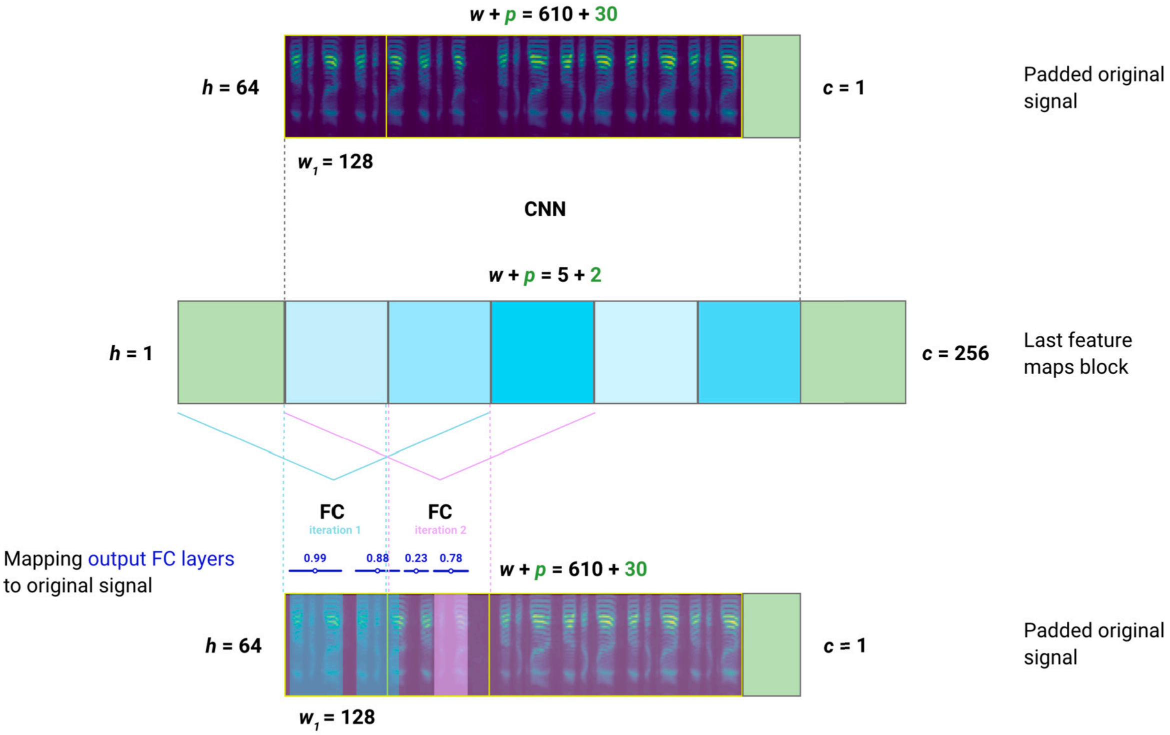


Fig. 8. DDK-AID network flow. The network takes in input the padded original spectrogram with dimension $(w+p,\ h,\ 1)$ where w are the original signal's frames (*i.e.,* resulting from the concatenation of the selected $Len(\tilde{f}_{i,j})$ and $Len(\tilde{e}_j)$), p is a padding to set $w+p=128k, k \in \mathbb{N}$ and h are the bands (*i.e.,* 64). This input is processed via convolutional layers and iterative-fully connected (FC) layers and outputs n spectrogram-segments (defined by their centre and width) according to the task (*i.e.,* repetition of the /pa/, /pa-ta-ka/ and /oo-ee/ syllables). The iterative action of the FC layers on the output produced by the convolutional ones is shown too. Contextually the figure displays the mapping of the output produced by the FC layers to the input (which is shown here only to improve the understanding of the methodology) to identify the spectrogram-segments. The outputs of the network are depicted in blue and are: (i) the width of the segment delimiting each syllable repetition and its midpoint and (ii) the level of confidence associated to each predicted portion (in the figure these confidences are 0.99, 0.88, 0.23, 0.78, respectively). To simplify image interpretation, we considered a $w=610$-frames signal.

progressively reduce initial $w+p$ and h while increasing the number of feature maps (c) up to the last convolutional layer whose output is a features block with dimension $(w+p)/128$, h and c equal to 1 and 256, respectively.

To handle feature blocks from the last convolutional layer of variable size, we implement FC-iterative layers. Indeed, as shown in Fig. 8, the flattening operation is carried out iteratively on equally-sized portions (size $(s)=3$) of the last padded-features block resulting in FC input layers of 768 ($c \cdot s = 768$). These sizes are the result of an iterative procedure that runs a window of amplitude 3 and stride 1 along the last padded features-block. At each iteration the output of the FC layers (namely the spectrogram-segments defined by their centre and width, as shown in Fig. 8) is mapped onto the corresponding region of the padded spectrogram in input to the DDK-AID network. This is done by considering the proportion between the width of the padded original signal and that of the FC layers-input.

As in YOLO networks [27], the outputs of the FC layers, given by each iteration are (i) the width of the segment delimiting each syllable repetition and its midpoint identified within the portion of the last padded-features block associated with the iteration, and a (ii) confidence associated with each predicted portion (see Fig. 8). To filter out predicted portions with an associated confidence too low, we tuned a minimum exclusion confidence-threshold ($minTh$) as done in [35]. The tuning was conducted considering a confidence threshold varying between 0.1 and 0.9.

The choice of not deriving the bounding-boxes coordinates, normally in output from the detection networks [36], but only the information necessary to get the amplitude of each $Len(\tilde{f}_{i,j})$-portion stems from the will of lowering computation. Indeed the main clinical need there, was the retrieval of the number of syllables repeated by the subject for each oral DDK task [6]. We derived this outcome by counting the number of predicted portions in output from the DDK-AID.

4. Experimental protocol

4.1. Dataset

The dataset used in this work was recorded by bulbar-onset ALS subjects who consecutively referred to Azienda Ospedaliero-Universitaria delle Marche (Italy) and healthy control subjects. The subjects involved were both males and females and had a similar age distribution. We excluded from the study subjects with tracheostomy, percutaneous endoscopic gastrostomy, cognitive impairment, without a caregiver and concomitant diseases that could interfere with communication skills or could affect life expectancy. To this goal, bulbar-onset subjects underwent assessment scales: (i) the ALS functional rating scale revised (ALS-FRS-R) to assess the severity of the disease by characterising features such as patient's motor, breathing and swallowing abilities (total score = 48, the lower the worst), (ii) the Montreal cognitive assessment (MoCA) to assess a cognitive impairment (total score = 30, the lower the worst) and (iii) the dysphagia outcome and severity scale (DOSS) to assess dysphagia's severity *i.e.,* the patient's difficulty to swallow foods or liquids (total score = 7, the lower the worst). Our bulbar-onset ALS subjects' characteristics per-scale are listed below:

- $33 \leq$ ALS-FRS-R ≤ 42
- $26 \leq$ MoCA ≤ 30
- $3 \leq$ DOSS ≤ 6

After approval of the study by the ethics committee, written informed consent was signed by each subject involved. After the obtainment of the registration permission we sent them the link to use the Homely Care web-application.

Each audio signal preprocessing was performed with the procedure described in Sec. 3.2.

The training and validation set were composed by all on-the-fly generated audio signals (see Fig. 4) with the procedure described in Sec. 3.3 and they differ from each other for the subjects involved, *i.e.,* 80% of the subjects were used to train while 20% to validate implementing a stratified-sampling fashion.

The test set had original audio recordings acquired by bulbar-onset ALS and healthy control subjects involved (see Fig. 4). Specifically, the bulbar-onset ALS data used to test the network derived from the 37 minutes of original audio recordings excluding the 21 minutes of recordings exploited for CNN training and validation (see Sec. 3.2). In a different manner, for the healthy control subjects we used the original audios from which the on-the-fly generated audios were derived.

In the training and validation sets each spectrogram-segment was noted (namely the width of the segment and its centre, see Fig. 7) along-side the number of repeated syllables. The annotation of each audio recording in the testing set, consisted in the number of repeated syllables, namely the actual outcome of clinical interest for the speech language pathologists involved in the study.

4.2. Training settings

To train the DDK-AID network, we set an initial learning rate of 0.0001. We set a number of epochs equal to 100 and used a batch size of 8 as a trade-off between memory constraints and computational resources available. Adam was used to optimize the network. The implemented loss function was inspired by [27] and was re-engineered to pursue the task of interest and save computation. In particular, all the YOLO inspired networks output 2D bounding boxes. Here we lower this output from 2D to 1D as to handle spectrogram-segments, as follows:

$$Loss = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(x_i + \hat{x}_i)^2] +$$

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})] + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(C_i + \hat{C}_i)^2] +$$

$$\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} [(C_i + \hat{C}_i)^2] \quad (8)$$

Where x_i identifies the actual location of the midpoint of the segment delimiting the syllable-repetition while $\hat{x}_i$ identifies the predicted coordinate. w_i represents the actual width of the segment delimiting the syllable-repetition while $\hat{w}_i$ is its predicted version. $\hat{C}_i$ represents the predicted confidence score of whether there is the syllable or not. λ_{coord} and λ_{noobj} are set as [27] as well as 1_{ij}^{obj} and 1_{ij}^{noobj}.

Inspired by YOLO X, the number of predictable portions in output from each feature block can be established in the DDK-AID network too. This number was experimentally set for each of the 3 tasks of interest considering the maximum number of each-syllable repetitions performed by a healthy subject in a single feature block (i.e., 2.048 s) from the last feature maps block. Particularly for the /pa/, /oo-ee/ and /pa-ta-ka/ syllable repetition oral-DDK task, this number was equal to 14, 10, 4, respectively.

All our analyses were performed using PyTorch framework on a Intel® Xeon® Silver 4214 CPU @ 2.20 GHz with 230 GB of RAM and a NVIDIA® RTX 2080 8 GB RAM.

4.3. Comparative method and evaluation metrics

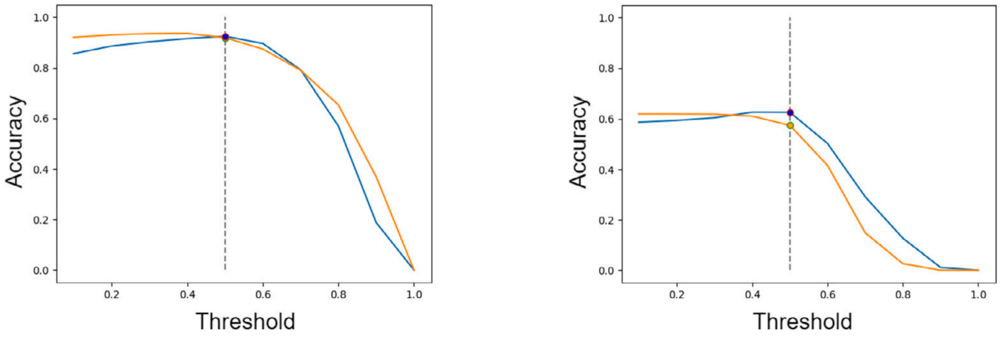
The performance of the DDK-AID network was compared against the one of the original YOLO X [21]. This CNN was the one which inspired our DDK-AID and, unlike [28], has a low computational complexity and, consequently, low cost for cloud-computing deployment. The YOLO X, unlike the DDK-AID, does not implement the FC-iterative layers but applies classical FC blocks to the entire flattened feature map in output from the last convolutional block. Therefore, the input signal of variable duration has been stretched in such a way as to allow the FC layers, after the flattening operation, to get equally-sized inputs.

For the YOLO X network the number of predicted portions in output from each feature block was experimentally set too. For the /pa/, /oo-ee/ and /pa-ta-ka/ syllables-repetition task, this number was set equal to 28, 20, 8, respectively.

For fair comparisons, the same training settings described in Sec. 4.2 were used as well as the same training/validation/testing set splits.

Both the networks performance were assessed in terms of Accuracy in pursuing the task of interest, *i.e.,* assessing the number of per-task repeated syllables, according to the following equation:

$$\text{Accuracy} = 1 - \frac{|\text{Syllables count}_{\text{actual}} - \text{Syllables count}_{\text{predicted}}|}{\text{Syllables count}_{\text{actual}}} \quad (9)$$



(a)

(b)

Fig. 9. Sample of confidence-threshold tuning for the /pa-ta-ka/ syllable repetition oral DDK task for the DDK-AID model (a) and YOLO X (b). In orange the curve for healthy control subjects and in blue the one for the bulbar-onset amyotrophic lateral sclerosis (ALS) subjects.

Table 2
Quantitative results in terms of Accuracy achieved by the two tested architectures, namely the DDK-AID and the YOLO X as well as the number of parameters (#Param.) for each architecture. Results are shown for both bulbar-onset amyotrophic lateral sclerosis (ALS) patients and healthy control subjects.

DDK-AID Network			
		Accuracy	
	#Param.	Bulbar-onset ALS patients	Healthy control subjects
/pa/	722986	**95.67**	**96.47**
/pa-ta-ka/	719116	**92.62**	**91.88**
/oo-ee/	721428	**82.57**	**95.45**

YOLO X			
		Accuracy	
	#Param.	Bulbar-onset ALS patients	Healthy control subjects
/pa/	1055304	42.92	52.32
/pa-ta-ka/	977904	57.94	61.72
/oo-ee/	1024344	45.93	57.39

Where the Syllables count$_{\text{predicted}}$ and the Syllables count$_{\text{actual}}$ represent the number of syllables predicted by the network and the ground truth provided by speech language pathologists, respectively.

5. Results

Fig. 9 shows the $minTh$ tuning outcome for the oral-DDK task of repeating the /pa-ta-ka/ syllables both for the proposed DDK-AID network and for the YOLO X. As visible from the results the highest performance was achieved for both the networks by a confidence threshold equal to 0.5. The same trend occurs for the other two oral-DDK tasks.

Table 2 shows the number of parameters of each tested architecture (*i.e.,* the DDK-AID network and the YOLO X one) for the /pa/, /pa-ta-ka/ and /oo-ee/ syllables repetition oral-DDK task. Considering the DDK-AID network, the architecture with lower number of parameters (Param.) was the one for assessing the number of syllables count during the /pa-ta-ka/ syllables-repetition task (# Param = 719116). This is followed by the architectures for evaluating subjects' performance while repeating the syllables /oo-ee/ (# Param = 721428) and /pa/ (# Param = 722986), respectively. The same trend can be seen for YOLO X network. This subtle difference in the number of trainable parameters, depends on the fact that, inspired by [21,27], we set the number of predicted portions in output from each feature block differently for each of the tasks. Particularly, for both the CNNs this number had the highest values for the /pa/ syllables repetition task ($= 14$, 28, for the DDK-AID and YOLO X, respectively) followed by the /oo-ee/ ($= 10$, 20, for the DDK-AID and YOLO X, respectively) and /pa-ta-ka/ ($= 4$, 8, for the DDK-AID and YOLO X, respectively) syllables-repetition tasks, respectively. In general, the DDK-AID network has fewer trainable parameters than YOLO-X.

The best Accuracy results are achieved by the DDK-AID network for both the ALS (mean Accuracy = 90.23) and healthy control subjects (*i.e.,* mean Accuracy = 94.60). For the YOLO X the mean accuracy for the bulbar-onset ALS subjects and healthy control subjects was equal to 48.93 and 57.14, respectively.

Observing the performance of the DDK-AID network, the worst results are achieved by the /oo-ee/ syllables-repetition task (Accuracy for bulbar-onset ALS subjects equal to 82.57). While, for the /pa/ and /pa-ta-ka/ syllables-repetition tasks, the same network achieves higher performance (Accuracy equal to 95.67 and 92.66, respectively). This performance reduction is also detectable in the scatterplots (Fig. 10). Each scatterplot shows the Accuracy of the architecture for the individual subjects involved in the study: bulbar-onset ALS subjects (blue dots) and healthy control subjects (orange dots). The X-axis of the graph shows the Syllables count$_{\text{actual}}$ and the Y-axis the Syllables count$_{\text{predicted}}$ by the DDK-AID (first row) and YOLO X (second row). The straight line in the graph distinguishes 3 regions: (i) above the straight line are the audio acquisitions for which the CNNs overestimates the Syllables count$_{\text{predicted}}$, (ii) below the straight line are the acquisitions for which the architectures underestimate the Syllables count$_{\text{predicted}}$, (iii) for all the predictions lying on the straight line the Syllables count$_{\text{predicted}}$ matches the Syllables count$_{\text{actual}}$. Observing the C. scatterplot (the one related to the /oo-ee/ syllables-repetition task) in Fig. 10, most of the errors (*i.e.,* the dots deviating from the line) concern the Syllables count$_{\text{predicted}}$ for the bulbar-onset ALS subjects. Even when viewing the training curves (Fig. 11) the one related to the /oo-ee/ syllables repetition oral-DDK task is that where the Loss on the validation deviates more from the one of the training set. Moreover, the scatterplots show that the /pa/ syllable repetition task is the one which induces less fatigability both in bulbar-onset ALS subjects and healthy control subjects compared to the polysyllabic tasks (*i.e.,* /pa-ta-ka/ and /oo-ee/ syllables repetition) [37]. In addition, from the graphs it appears that healthy control subjects mainly tend to do more syllables repetitions per task than bulbar-onset ALS subjects as stated in [23,12]. These results confirmed that bulbar-onset ALS subjects – who develop impairments in the orofacial musculature – progressively show slower articulation of words [38]. From the graph is visible that YOLO-X, unlike our DDK-AID, tends to underestimate the Syllables count$_{\text{predicted}}$ for each of the task. Though both the CNNs tend to overestimate critically ill subjects (*i.e.,* those with lower values of Syllables count$_{\text{actual}}$).

6. Discussion

The proposed work presents a support system based on deep learning to automatically and remotely assess a subject while performing the oral-DDK test. The system consists of a web application to acquire audio recordings while repeating the /pa/, /pa-ta-ka/ and /oo-ee/ syl-

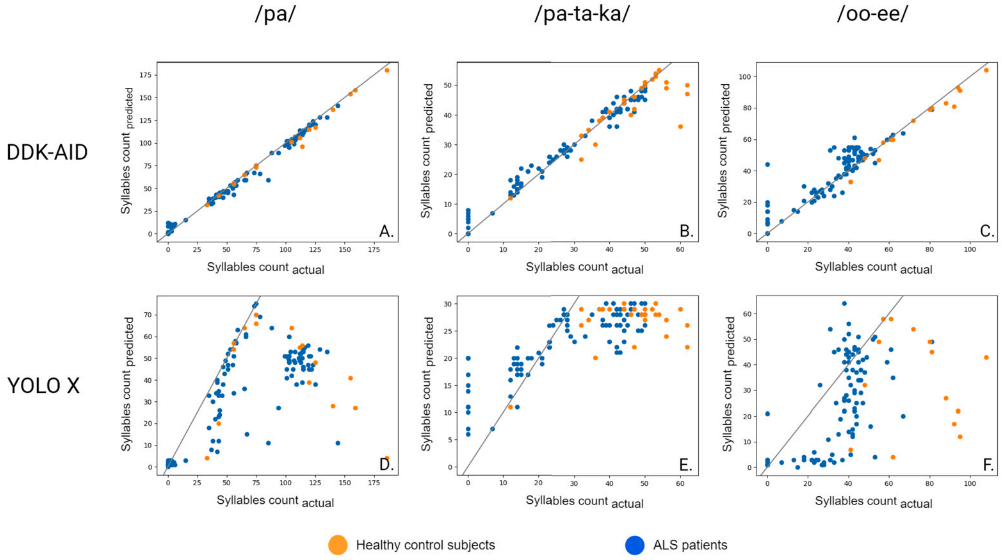


Fig. 10. Scatter plots representing the variable Syllables count$_{actual}$ on the x-axis and the variable Syllables count$_{predicted}$ on the y-axis. The first, second and third columns show the scatterplots for the /pa/, /pa-ta-ka/ and /oo-ee/ oral-DDK task, respectively. The first row shows the scatterplots of the DDK-AID while the second one those of YOLO-X.

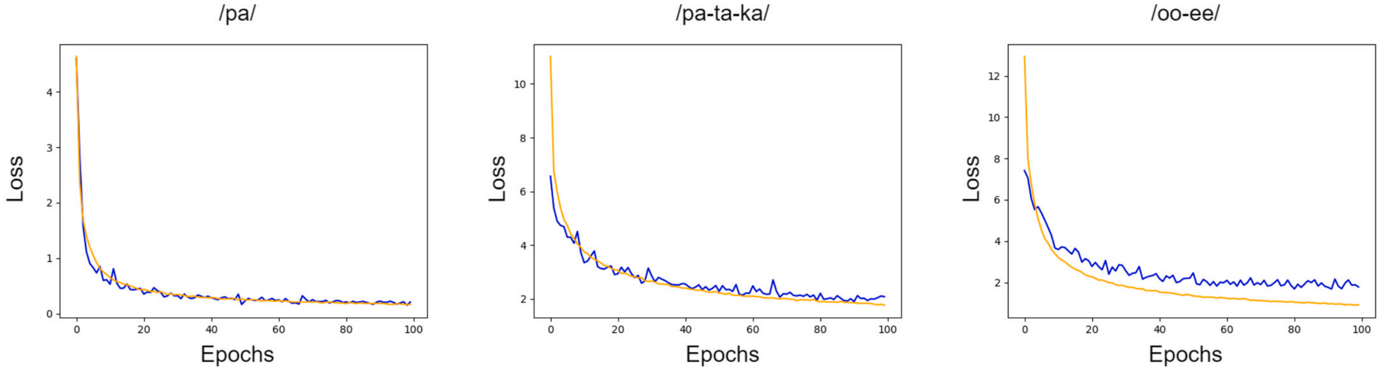


Fig. 11. Training curves of the DDK-AID network for the 3 oral-DDK tasks (*i.e.,* /pa/, /pa-ta-ka/, /oo-ee/ syllables repetition). In orange the training Loss and in blue the validation one.

lables. These recordings are used to train and test the DDK-AID network which outputs the number of per-task repeated syllables. The network is specifically designed to process audio recordings of variable duration. Moreover, it handles the scarcity and variability of data collected through an on-the-fly training-data generation based on a synthetic-signals generation algorithm.

As showed in Sec. 5, the proposed DDK-AID achieves improved performance with respect to the YOLO X. Indeed, the YOLO X CNN, unlike the DDK-AID one, is not designed to process audio acquisitions of different duration. Therefore, in the preprocessing step, the audio-signal stretching is required so that the FC layers get equally sized inputs. However, this resizing suffers from main implications: (i) the original signal degrades especially when its length deviates excessively from that required as input by the CNN and (ii) the additional stretching phase introduces a variability factor in the training data which may harden the task.

Our DDK-AID has fewer parameters than the YOLO X. Indeed, in the DDK-AID network there is an exact relationship between signal frames and time. This allows a maximum number of predictable portions per feature block to be precisely set on the basis of the maximum number of syllables repetitions physically achievable by the healthy control group. This relationship is lost when dealing with signals to which a stretching factor is applied. The latter aspect forces us to define, after proper tuning, a maximum number of repetitions in excess for the task of our interest.

It should also be emphasised that the iterative-FC layers in our DDK-AID, unlike those of the YOLO X which are classical FC, allow to process data streams of any length without performance degradation. Indeed, these layers enable our CNN to count the number of syllable repetitions in the initial portions of the signal while continuing to receive and process data from the stream as to predict the total number of syllables repeated by the subject.

The proposed training technique, based on audio-signals generation (Sec. 3.3), introduced two limitations visible in Sec. 5: the reduced performance (i) in the /oo-ee/ syllables-repetition task for bulbar-onset ALS subjects, (ii) on critically ill subjects (mainly visible in the scatterplots B. and C. of Fig. 10). Regarding the issue (i) in the original signal, we experimentally verify that the representation in the spectrogram of the current syllable /oo-ee/ shows a dependency with the adjacent ones (*i.e.,* previous and subsequent syllable repetition). The concatenation between individual syllables implemented with our data-augmentation technique, does not consider the syllables dependency, producing less-realistic signals. Issue (ii) is caused by the fact that all the generated signals are obtained by concatenating, in an alternating way, portions with syllable repetition and silence portions. This pattern however is no more valid in the audio acquired from critically-ill bulbar-onset ALS subjects. Indeed, mainly due to mispronunciation, not all the syllable-repetition portions are annotated as an actual repetition by the speech language pathologists.

A straightforward limitation of the implemented methodology can be seen in the way we partitioned the data for training and testing, as well as in the choice of generating audio signals from real ones for training purposes. However, our aim is to investigate the feasibility of proposing a deep-learning method, integrable into a telemedicine system, to (i) analyse data collected in an uncontrolled scenario and with commonly used devices, (ii) process audio acquisitions of different duration in line with the clinical needs for conducting the evaluation [31]. To the best of our knowledge, this work is among the first to investigate these latter aspects.

As future work, to mitigate the aforementioned issues, we are going to increase our dataset size and variability and to make our data publicly available for promoting research in this field. Furthermore, we plan to improve the on-the-fly training-data generation mechanism by chaining together the actual syllables repetitions, silences and incorrectly pronounced repetitions and to couple the generated signals with real ones during training.

7. Conclusion

The results obtained from our deep-learning-driven system designed to aid clinicians in evaluating the evolution of dysarthria in subjects with bulbar-onset ALS are promising. Nevertheless, we recognize that further investigation and validation are essential to progress toward an improved framework for enhancing the well-being of these patients.

In this perspective, our future work will be devoted to the study of other characteristics that can be acquired from the audio recorded during the diadochokinesis task (*e.g.,* the length of pauses between two successive syllables). Moreover, all the algorithms for the analysis of the audio-recordings will be integrated within a broader telemonitoring system that also includes the assessment of orofacial functions related to speech [17]. Ultimately, system usability will be soon investigated through structured and semi-structured interviews with domain experts involving both patients and their caregivers.

Statement of ethical approval

The study we conducted fully respects and promotes the values of freedom, autonomy, integrity and dignity of the person, social solidarity and justice, including fairness of access. The study was carried out in compliance with the principles laid down in the Declaration of Helsinki, in accordance with the Guidelines for Good Clinical, after obtaining the approval of the Ethics Committee of the *Azienda Ospedaliero-Universitaria delle Marche di Ancona*, Italy (Protocol-ID 118 25/3/2021).

Declaration of competing interest

• All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

• This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

• 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.

Acknowledgement

This work is part of the *Parola di Motoneurone* project which won the prize of the Italian Society of Neurology 2020 call.

References

[1] E. Longinetti, F. Fang, Epidemiology of amyotrophic lateral sclerosis: an update of recent literature, Curr. Opin. Neurol. 32 (5) (2019) 771.
[2] A. Chiò, G. Mora, A. Calvo, L. Mazzini, E. Bottacchi, R. Mutani, et al., Epidemiology of als in Italy: a 10-year prospective population-based study, Neurology 72 (8) (2009) 725–731.
[3] L.P. Rowland, N.A. Shneider, Amyotrophic lateral sclerosis, N. Engl. J. Med. 344 (22) (2001) 1688–1700.
[4] J.R. Duffy, Motor Speech Disorders e-Book: Substrates, Differential Diagnosis, and Management, Elsevier Health Sciences, 2019.
[5] C. Barnett, J.R. Green, R. Marzouqah, K.L. Stipancic, J.D. Berry, L. Korngut, A. Genge, C. Shoesmith, H. Briemberg, A. Abrahao, et al., Reliability and validity of speech & pause measures during passage reading in ALS, Amyotroph. Lateral Scler. Frontotemporal Degener. 21 (1–2) (2020) 42–50.
[6] P. Rong, Automated acoustic analysis of oral diadochokinesis to assess bulbar motor involvement in amyotrophic lateral sclerosis, J. Speech Lang. Hear. Res. 63 (1) (2020) 59–73.
[7] S.J. Robertson, Dysarthria Profile, Communication Skill Builders, 1987.
[8] P. Rong, Y. Yunusova, J. Wang, L. Zinman, G.L. Pattee, J.D. Berry, B. Perry, J.R. Green, Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems, PLoS ONE 11 (5) (2016) e0154971.
[9] A. Eisen, M. Kiernan, H. Mitsumoto, M. Swash, Amyotrophic lateral sclerosis: a long preclinical period?, J. Neurol. Neurosurg. Psychiatry 85 (11) (2014) 1232–1238.
[10] G.L. Pattee, E.K. Plowman, K.L. Garand, J. Costello, B.R. Brooks, J.D. Berry, R.A. Smith, N. Atassi, J.L. Chapin, Y. Yunusova, et al., Provisional best practices guidelines for the evaluation of bulbar dysfunction in amyotrophic lateral sclerosis, Muscle Nerve 59 (5) (2019) 531–536.
[11] S. Roldan-Vasco, A. Orozco-Duque, J.C. Suarez-Escudero, J.R. Orozco-Arroyave, Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia, Comput. Methods Programs Biomed. 208 (2021) 106248.
[12] H. Ackermann, I. Hertrich, T. Hehr, Oral diadochokinesis in neurological dysarthrias, Folia Phoniatr. Logop. 47 (1) (1995) 15–23.
[13] A. Rueda, J.C. Vásquez-Correa, J.R. Orozco-Arroyave, E. Nöth, S. Krishnan, Empirical mode decomposition articulation feature extraction on Parkinson's diadochokinesia, Comput. Speech Lang. 72 (2022) 101322.
[14] L.J. Ball, D.R. Beukelman, G.L. Pattee, Timing of speech deterioration in people with amyotrophic lateral sclerosis, J. Med. Speech Lang. Pathol. 10 (4) (2002) 231–236.
[15] J. Lee, A. Madhavan, E. Krajewski, S. Lingenfelter, Assessment of dysarthria and dysphagia in patients with amyotrophic lateral sclerosis: review of the current evidence, Muscle Nerve 64 (5) (2021) 520–531.
[16] J.S. Hong, C. Wasden, D.H. Han, Introduction of digital therapeutics, Comput. Methods Programs Biomed. 209 (2021) 106319.
[17] L. Migliorelli, D. Berardini, K. Cela, M. Coccia, L. Villani, E. Frontoni, S. Moccia, A store-and-forward cloud-based telemonitoring system for automatic assessing dysarthria evolution in neurological diseases from video-recording analysis, Comput. Biol. Med. (2023) 107194.
[18] J.R. Green, Y. Yunusova, M.S. Kuruvilla, J. Wang, G.L. Pattee, L. Synhorst, L. Zinman, J.D. Berry, Bulbar and speech motor assessment in als: challenges and future directions, Amyotroph. Lateral Scler. Frontotemporal Degener. 14 (7–8) (2013) 494–500.
[19] Y.Y. Wang, K. Gao, A.M. Kloepper, Y. Zhao, M. Kuruvilla-Dugdale, T.E. Lever, F. Bunyak, Deepddk: a deep learning based oral-diadochokinesis analysis software, in: 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE, 2019, pp. 1–4.
[20] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[21] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: exceeding yolo series in 2021, arXiv preprint, arXiv:2107.08430, 2021.
[22] M. Novotnỳ, J. Rusz, R. Čmejla, E. Ržička, Automatic evaluation of articulatory disorders in Parkinson's disease, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (9) (2014) 1366–1378.
[23] M. Novotny, J. Melechovsky, K. Rozenstoks, T. Tykalova, P. Kryze, M. Kanok, J. Klempir, J. Rusz, Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis, J. Speech Lang. Hear. Res. 63 (10) (2020) 3453–3460.
[24] Y. Bengio, Y. Lecun, G. Hinton, Deep learning for AI, Commun. ACM 64 (7) (2021) 58–65.
[25] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[26] M. Müller, M. Gromicho, M. de Carvalho, S.C. Madeira, Explainable models of disease progression in als: learning from longitudinal clinical data with recurrent neural networks and deep model explanation, Comput. Methods Programs Biomed. Update 1 (2021) 100018.
[27] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[28] K. Rozenstoks, M. Novotny, D. Horakova, J. Rusz, Automated assessment of oral diadochokinesis in multiple sclerosis using a neural network approach: effect of different syllable repetition paradigms, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (1) (2020) 32–41, https://doi.org/10.1109/TNSRE.2019.2943064.

[29] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization 64 (3) (2021) 107–115, https://doi.org/10.1145/3446776.

[30] X. Wang, Y. Han, V.C. Leung, D. Niyato, X. Yan, X. Chen, Convergence of edge computing and deep learning: a comprehensive survey, IEEE Commun. Surv. Tutor. 22 (2) (2020) 869–904.

[31] C. Fussi, Profilo di valutazione della disartria, adattamento italiano del test di Robertson, raccolta di dati normativi e linee di intervento, 2010.

[32] P. Zhang, X. Zhang, Deep template matching for small-footprint and configurable keyword spotting, in: INTERSPEECH, 2020, pp. 2572–2576.

[33] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid, et al., Automatic speech analysis to early detect functional cognitive decline in elderly population, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 212–216.

[34] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intell. Data Anal. 6 (5) (2002) 429–449.

[35] B. Jiang, R. Luo, J. Mao, T. Xiao, Y. Jiang, Acquisition of localization confidence for accurate object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 784–799.

[36] S. Ren, K. He, R. Girshick, X. Zhang, J. Sun, Object detection networks on convolutional feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 39 (7) (2016) 1476–1481.

[37] B. Peter, H. Lancaster, C. Vose, K. Middleton, C. Stoel-Gammon, Sequential processing deficit as a shared persisting biomarker in dyslexia and childhood apraxia of speech, Clin. Linguist. Phon. 32 (4) (2018) 316–346.

[38] H. Zhang, L. Chen, J. Tian, D. Fan, Disease duration of progression is helpful in identifying isolated bulbar palsy of amyotrophic lateral sclerosis, BMC Neurol. 21 (1) (2021) 1–8.