IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Artificial intelligence for dysarthria assessment in children with ataxia: a hierarchical approach

G. Tartarisco[1], R. Bruschetta[1], S. Summa[2], Member, IEEE, L. Ruta[1], M. Favetta[2], M. Busà[1], A. Romano[2], E. Castelli[2], F. Marino[1], A. Cerasa[1], T. Schirinzi[4], M. Petrarca[2], E. Bertini[3], G. Vasco[2], G. Pioggia[1]

[1]Institute for Biomedical Research and Innovation (IRIB-CNR), National Research Council of Italy (CNR), 98164 Messina, Italy
[2]MARlab, Neurorehabilitation Division, Intensive Neurorehabilitation and Robotics Department, IRCCS Bambino Gesù Children's Hospital, IRCCS, Rome, Italy
[3]Unit of Neuromuscolar and Neurodegenerative Diseases, Department of Neurosciences, IRCCS Bambino Gesù Children's Hospital, IRCCS, Rome, Italy
[4]Department Systems Medicine, University of Roma Tor Vergata, Rome, Italy

Corresponding author: Gennaro Tartarisco (gennaro.tartarisco@cnr.it)

(Gennaro Tartarisco and Roberta Bruschetta are co-first authors)
(Giovanni Pioggia and Gessica Vasco are equally senior authors)

The study was supported by Progetto di Rete NET-2013-02356160-3, Italian Ministry of Health.

ABSTRACT Early onset ataxia represents a group of heterogeneous neurological conditions typically characterized by motor disability. Speech problems are one of the core features of ataxic syndromes; hence, the automatic characterization of speech impairment may represent a source of biomarkers for early screening and stratification of patients. The main contribution of this paper consists in proposing a novel hierarchical machine learning model (HMLM) to improve detection and assessment of dysarthria from a structured speech disturbance test. Performances are tested on a new audio dataset containing 10 seconds recordings of standardized clinical PATA test for 55 subjects: 18 healthy subjects and 37 with ataxia. Results show that the proposed HMLM achieves performances with an accuracy of about 90% at the first level (healthy vs patients) selecting an optimal subset of conventional features. In cascade, at the second level, speech disturbance severity (Low vs High) is assessed using deep learning feature extraction technique based on a VGG pre-trained network with maximum accuracy of about 80%. Both levels are processed through the majority voting ensemble technique testing Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree (DT) and Naïve Bayes (NB). In our results, the use of HMLM considerably outperforms the results achieved with a single machine learning or deep learning modeling. These outcomes demonstrate that the investigation of the PATA speech test through HMLM can be considered very promising. We also observed that the use of conventional feature extraction techniques and machine learning modeling seems to be a good solution for the diagnosis of patients with ataxia, while the deep learning approach is more appropriate for stratification of severity of dysarthria.

INDEX TERMS ataxia, dysarthria, speech disturbance, pata test, deep learning, feature extraction, hierarchical systems, machine learning, speech recognition.

I. INTRODUCTION

Early onset ataxia (EOA) represents a heterogeneous group of neurological disorders, with inherited or acquired aetiology and usually with onset before 25 years [1]. Depending on the clinical progression, they can be divided in progressive ataxias (PAs) and congenital non progressive ataxias (CAs), including respectively entities with different aetiology, phenomenology, and prognosis [1]–[7]. Regardless of such clinical features, although rare (estimated European prevalence 26/100,000) [8], EOAs are responsible for relevant disability and high

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

costs, since no effective treatment is still available [9]–[11]. Patients suffer many neurological disturbances responsible for severe physical limitations, which negatively affect their well-being [9]. Generally, ataxia is characterized by coordination disturbances with an effect on walking, standing and voluntary movements of the upper limb. Moreover, patients can manifest speech disturbances that can be responsible for communicative and social limitations, significantly decreasing patients' quality of life [9]. Indeed, dysarthria (motor difficulties in speech) and dysphagia (motor difficulties in swallowing) are frequent signs in ataxic syndromes.

The number of experimental trials, covering both potential disease-modifying treatments [12] and symptomatic interventions (physical therapy or neuromodulation) [13], are significantly increasing in the ataxia field. Indeed, there is an urgent need of specific and reliable biomarkers, either for early stratification of patients or for the accurate monitoring and follow-up. Actually, assessment of patients with ataxia currently relies on the clinical scores, such as the Scale for the Assessment and Rating of Ataxia (SARA) [14]. In the case of speech, it can be assessed by the perceptual tests, where an expert listener rates 21 parameters of speech considering prosody, respiration, phonation, resonance, intelligibility, naturalness and articulation. A complete evaluation of dysarthria in Friedreich's ataxia has been reported by Folker et al., in 2010 [15]. Some limitations of clinically-based ataxia rating methods are: rater variability, the ceiling and floor effects [1], [16], [17] and the loss of accuracy, particularly in the pediatric age [17], [18]. In the last years, the use of technologies is providing a promising help showing reliable, objective, accurate and continuous outcome measures either in conventional and "telemedicine" settings [19]–[26]. An objective assessment of speech could represent a potential source of biomarkers. Indeed, it has been proven, in several neurodegenerative diseases, that there is a relationship between oral motor deficits and CNS integrity [15], [27]–[29]. Concordantly, objective measures of speech have been suggested as meaningful information on the patient and health-related quality of life in clinical trials [30], [31]. However, research still lacks natural history studies of speech disturbances in patients with ataxia [15], [31], [32]. As already suggested in [35], to assess dysarthria in children with ataxia, artificial intelligence has shown very promising results. A fundamental step is the extraction of all features that can be used as input parameters in disorder characterization systems [33], [34]. The aim is to identify the relevant information contained in the speech signal. Binary classifiers are commonly used to distinguish pathological from the healthy condition [35], [36]. For instance, Rudzicz et al. [35], employed feed-forwards artificial neural networks (ANNs), and SVMs with phonological features have been used to design discriminative models for dysarthric speech. A binary classifier [37], based on Mahalanobis distance and discriminant analysis was developed for dysarthria severity classification, where 95% accuracy was achieved. An automatic intelligibility assessment system that performs a binary classification by capturing atypical variation in dysarthric speech by using linear discriminant analysis (LDA), k-nearest neighbor (KNN) and SVM classifiers was proposed [36] with an accuracy of 68%, 66% and 70% respectively. While in [33] four levels of intelligibility were recognized with an accuracy between 40-50%, using SVMs and testing different feature sets. Moreover, the combination of the statistical GMM and ANNs was used in [38], achieving accuracy of 86% over three degrees of severity levels. Speaker identification (97.2%) and severity level assessment (93.2%) revealed the best performance using SVMs and hybrid GMM/SVM systems in [34]. Existing studies were carried out through the employment of the few available dysarthric speech databases such as TORGO [39] and NEMOURS [40]. Both of these databases include few subjects (not more than 15) with different levels of dysarthria, due to various conditions such as cerebral palsy (CP), head trauma (HT) and amyotrophic lateral sclerosis (ALS). They are composed of short sentences and words or acoustic and articulatory features extracted from them. The lack of suitable and sufficient data is one of the biggest limits in the field of analysis of speech and verbal communication disorders. Moreover, in our specific case of ataxia, the design and collection of a suitable database is a critical issue since it is a rare genetic group of disorders and there are constraints such as recording conditions, patient's availability, and approval of health agencies.

Here we developed a tool aimed at automatically recognizing ataxic syndromes. For this purpose, we collected recordings of their speech disturbance assessment, made through the standardized clinical PATA speech test of the SARA scale. To our knowledge, this is the first study dealing with artificial intelligence for the assessment and stratification of severity of dysarthria in ataxia through a standardized clinical speech test. In our case, we developed a novel HMLM based on a fusion of conventional and deep learning features to automatically assess the healthy vs patients and quantify the level of speech disturbance. Results demonstrate that the use of two binary models of artificial intelligence in cascade, outperforms compared to a single machine learning or deep learning classifier. The results obtained are encouraging and highlight the validity of HMLM with mixed conventional and deep learning features to recognize ataxia and stratify the level of severity of dysarthria. However, an extensive validation phase on a greater number of subjects is needed. In fact, we plan to continue to test a higher number of subjects to validate the HMLM applied to the PATA speech test as a tool to support clinicians for optimizing screening, clinical tests and personalized treatments.

II. METHODS

A. OVERALL ARCHITECTURE

The HMLM model is the main component developed and tested for the assessment of ataxia. Given pre-processed "PA-TA" speech data, the first level of machine learning (ML) processes and discriminates healthy vs patients. Once the speech disease is detected the second level of ML assesses the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

severity of dysarthria. The overall system is detailed in Fig. 1, which shows the data flow and indicates which features and ML are selected and tested for each level to achieve the best performance. All these elements are detailed in the following sections.

B. DATA COLLECTION

1)PATIENTS ENROLLMENT

The study population was recruited in 2018 at the Movement Analysis and Robotics laboratory (MARlab) of the Intensive Neurorehabilitation and Robotics Departments of IRCCS Bambino Gesù Children's Hospital (Rome, Italy). Overall, it is composed of 55 subjects: 18 healthy (H), 21 with Progressive Ataxia (PA) and 16 with Congenital non Progressive Ataxia (CA). H group included sex/age-matched healthy volunteers without personal/familiar history of neurological diseases and no signs at clinical examination (age 12[7.6]; 12F/6M). All patients had genetically confirmed diagnosis and a routine diagnostic workup, including general and neurological examination, brain MRI, sensory evoked potentials, nerve conduction study and visual acuity evaluation; moreover, they were in follow-up at the MARlab for at least 2 years, to ensure a correct group classification. None of the enrolled subjects had relevant cognitive impairment or were taking psychoactive drugs (other usual medications, such as vitamin or antioxidant were allowed). Patients with severe disability, moderate-severe cognitive impairment affecting tests execution were excluded. Demographic data were collected for the three groups. The research conformed to the ethical standards laid down in the 1964 Declaration of Helsinki. All subjects participated on a voluntary basis, after that they or their legal responsible signed the informed consent (the study was approved by local ethical committee Protocol NET-2013-02356160 WP3, nr. 1619-2018, received 03 July 2018).

2)EXPERIMENTAL SETUP

After receiving a clinical evaluation, all the 55 subjects were asked to perform the "PATA" test in a quiet room. Each vocal task was recorded with SaraHome, a novel technology for the assessment at home of patients with ataxia symptoms [41], using the microphone array mounted on the Microsoft Kinect V2 for 10 seconds at sampling frequency (F_s) of 16 KHz. Each subject was asked to repeat the word "PATA" as many times as possible in 10 seconds, as reported in [42], [43]. At the end of each task, speech disturbance was scored by expert personnel using a standardized clinical scale: SARA [14]. For each patient with CA and PA, the same test was repeated after 12 months (time t_1) to monitor the possible evolution of disturbances. It was possible to repeat the test only for 21/34 patients (12 PA and 9 CA) For this reason, 76 audio recordings were totally considered. All the data were analyzed using Matlab version 2020 (Mathworks, Natick MA).

C. SIGNAL PRE-PROCESSING AND "PA-TA" SEGMENTATION

Sometimes the collected data were affected by background noise such as external voices, door slamming sounds or environmental noises; therefore, a step of pre-processing and clean-up was necessary. Initially, we evaluated the average signal spectrum (Short Time Fourier periodogram) to detect the frequency range of interest (Fig. 2). Since patients' voice repeating "PA-TA" was mostly under the frequency of 1 kHz, in order to reduce all the noise above this frequency, we applied an eleven-order low-pass Chebishev filter with a cut-off frequency of 1 kHz and a Hanning window with length equal to the 0.5% of F_s. After, we applied the method based on fine-tuning of threshold short-term energy and spectral spread [44] to detect speech boundaries and remove the remaining noise. The envelope of each signal was extracted by applying firstly the module of the Hilbert Transform and later a zero-phase moving-average filter whose parameters were tuned according to signal approximate entropy. If signal approximate entropy was lower than the empirical threshold of 0.8, a single moving-average filter was applied to the Hilbert Transform; otherwise, the two cascade filters were employed as reported in Table I. The main steps of signals pre-processing are shown in Fig. 3. After these steps, "PA" & "TA" peaks were detected from the envelope selecting only maxima with a minimum prominence equal to the 10% of the absolute value of Hilbert Transform mean. Instead, signal minima were recognized by computing the energy and by choosing only the minimum prominence of 0.01 and at least 10% of the sampling frequency apart.

D. FEATURE EXTRACTION AND SELECTION FOR MACHINE LEARNING

Audio signals were segmented in order to increase the statistical significance of the dataset in terms of inter-subject and inter-class variability [45]. Because of windowing the signals, it was possible to assume their quasi-stationary within each frame, easing the subsequent analysis [46]. Since the performance of the system depends largely on noise reduction among peaks and the selection of useful acoustic events only (Fig. 4), it was necessary to carry out the segmentation using the "PA" & "TA" peaks as reference points, and the samples between the closest preceding and consecutive minima considering each PA-TA cycle. After performing audio segmentation, we investigated the most relevant conventional features of our targeted application. In literature, the issue of feature extraction in the field of audio processing is quite challenging because of several factors such as the simultaneous presence of different sound sources and the background noises that may affect machines performance [46]. These characteristics are considered to identify the most reliable parameters. In Fig. 1 all the features extracted and grouped by time domain ($PATA_{freq}$, Approximate Entropy), frequency domain (spectral values, mfcc and gtcc coefficients), chaotic domain (Lyapunov Exponent) and Age of children, are listed. All the features were extracted from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

each PA-TA cycle and then the average value for each subject was calculated.

PATA frequency ($PATA_{freq}$), is a simple time domain physical feature, whose calculation is directly performed from the temporal envelope of signals in order to assess a fundamental parameter of our specific task, according to the following equation:

$$PATA_{freq} = \frac{n_{peaks}}{2 \cdot l} \quad (1)$$

Where n_{peaks} is the total number of recognized peaks and l is the length of the signal.

Approximate Entropy is calculated to measure the complexity and possible fluctuations of the signals [47] and for its strength to discriminate human voice components from corrupted speech [48]. Lyapunov Exponent is calculated to consider the non-linearity of speech [49]–[53]. Frequency domain features, conventionally used for lots of applications [54]-[55] are the most described in literature. These variables, are intended to describe the physical properties of the signal frequency content and they cover a large number of different categories. Among this wide range of possibilities, we computed the following features:

- *Mel-Frequency Cepstral Coefficients (MFCCs)*: They are one of the most popular features employed in speech processing. They constitute the mel-frequency cepstrum (*MFC*), a compact representation of the short-term power spectrum of an audio signal, obtained through a linear cosine transform from the log power spectrum to the nonlinear mel scale frequency [56].
- *Gammatone Cepstral Coefficients (GTCCs)*: They are a modification of MFCCs inspired from biology and are obtained applying Gammatone filters with equivalent rectangular bandwidth bands [57].
- *Spectral Centroid*: It can be considered the barycenter of the spectrum and indicates where most of signal energy is contained:

$$Spectral\ Centroid = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k} \quad (2)$$

Where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral centroid [55], [58].

- *Spectral Spread*: It is a measure of the spread of the spectrum around its mean value:

$$Spectral\ Spread = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 s_k}{\sum_{k=b_1}^{b_2} s_k}} \quad (3)$$

Where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral spread and μ_1 is the spectral centroid [56], [59].

- *Spectral Skewness*: It is a measure of the asymmetry of the spectrum around its mean value and is computed from the 3rd order moment:

$$Spectral\ Skewness = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k}{(\mu_2)^3 \sum_{k=b_1}^{b_2} s_k} \quad (4)$$

Where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral skewness, μ_1 is the spectral centroid and μ_2 is the spectral spread.

Skewness=0 symmetric distribution
Skewness <0 more energy on the right
Skewness >0 more energy on the left [60]

- *Spectral Kurtosis*: It gives a measure of the flatness of the spectrum around its mean value and indicates a possible nonstationary or non-Gaussian behavior in the frequency domain. It is the 4th order moment and is computed starting from the short-time Fourier Transform of the signal *S(t,f)*:

$$Spectral\ Kurtosis = \frac{\langle |S(t,f)|^4 \rangle}{\langle |S(t,f)|^2 \rangle^2} - 2, f \neq 0 \quad (5)$$

Where $\langle \cdot \rangle$ is the time-average operator
Kurtosis=3 normal distribution
Kurtosis <3 flatter distribution
Kurtosis >3 peaker distribution [61]–[63]

- *Spectral Slope*: It represents the amount of decrease of the spectral amplitude and is computed as the linear regression of the spectral amplitude:

$$Spectral\ Slope = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_f)(S_k - \mu_s)}{\sum_{k=b_1}^{b_2} (f_k - \mu_f)^2} \quad (6)$$

Where f_k is the frequency in Hz and s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral slope, μ_f is the mean frequency and μ_s is the mean spectral value [64].

- *Spectral Decrease*: It represents the amount of decrease of spectral amplitude too, but it was defined from the perceptual studies to be more correlated to human perception.

$$Spectral\ Decrease = \frac{\sum_{k=b_1+1}^{b_2} \frac{s_k - s_{b_1}}{k-1}}{\sum_{k=b_1+1}^{b_2} s_k} \quad (7)$$

Where s_k is the spectral value that correspond to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral decrease.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

- *Spectral RolloffPoint*: It is the frequency below which there is 95% of the signal energy:

$$Spectral\ RolloffPoint = i \text{ so that } \sum_{k=b_1}^{i} s_k\ 0.95\ \sum_{k=b_1}^{b_2} s_k \quad (8)$$

Where s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral rolloffpoint.[65]

- *Spectral Flatness*: It is a measure of the noisiness/sinusoidality of a spectrum and is computed as the ratio between the geometric mean and the arithmetic mean of the energy spectrum:

$$Spectral\ Flatness = \frac{(\prod_{k=b_1}^{b_2} s_k)^{\frac{1}{b_2-b_1}}}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} s_k} \quad (9)$$

Where s_k is the spectral value that corresponds to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral flatness.
For tonal signals it is close to 0, for noisy signals it is close to 1 [66].

- *Spectral Crest*: It is a measure of the noisiness/sinusoidality of a spectrum too but it is computed as the ratio between the minimum value within the band and the arithmetic mean of the energy spectrum:

$$Spectral\ Crest = \frac{\max(s_{k\in[b_1,b_2]})}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} s_k} \quad (10)$$

Where s_k is the spectral value that correspond to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral crest.

- *Spectral Entropy*: It describes the complexity of the distribution:

$$Spectral\ Entropy = \frac{-\sum_{k=b_1}^{b_2} s_k \log(s_k)}{\log(b_2-b_1)} \quad (11)$$

Where s_k is the spectral value that correspond to bin k, while b_1 and b_2 are the band edges, in bins, over which to calculate the spectral entropy.[67]

- *Pitch*: It is the fundamental frequency of the audio signal, so that its integer multiple best explain the content of the signal spectrum [68]–[71].

- *Harmonic Ratio (HR)*: It is the ratio between the power of the fundamental frequency and the total power in an audio frame:

$$HR = \frac{\sum_{n=1}^{N} s(n)s(n-m)}{\sqrt{\sum_{n=1}^{N} s(n)^2 \sum_{n=0}^{N} s(n-m)^2}}\ 1 \le m \le M \quad (12)$$

Where s is a single frame of audio data with N elements and M is the maximum lag in the calculation [72], [73].

Once features have been extracted, the next step was to eliminate redundant variables preserving the amount of information and increasing computational speed and performances [74], [75]. Among the highly correlated variables (Spearman correlation ≥ 75% [74], [75]), the least correlated variables with the output of classification were removed as shown in Fig. 5, and features min-max normalization was implemented. After this step, the ranking of the univariate features according to the predictor importance score, was performed using chi-square tests [76]–[79]. Then the optimal subset of features was defined selecting the highest difference between consecutive scores as the break-point. Finally, the best combination of features was achieved by selecting 6 main features (mfcc3, Age, $PATA_{freq}$, Spectral Centroid, Spectral Kurtosis, mfcc8) as shown in Fig. 6. We tested different techniques for feature selection obtaining comparable results so that we chose the best feature selection technique in terms of the computational cost.

E. FEATURE EXTRACTION AND SELECTION WITH DEEP LEARNING

Deep Learning Networks are complex architectures used to detect specific features directly from data. They can have hundreds of layers and a huge number of parameters such as weights and bias to be learned. Training from scratch a deep architecture in order to extract specific features avoiding overfitting, requires a large amount of data (hundreds, thousands or even millions, it depends on the application) resulting in high computational and timing costs. Generally, the time of training is related to lots of different factors like the number of epochs, dataset size, computational power etc., but to reach a certain accuracy even months could be necessary. Usually, GPUs are employed to speed up the process. In many real applications, it is difficult and expensive to obtain training data that match the feature space and predict the distribution characteristics of the test data. Therefore, in practice there is a need to create a high-performance learner for a target domain trained from a related source domain. This is the motivation for the transfer learning [80]. Leveraging a pretrained network that has already learned many features on a big dataset to exploit it for a new task, and to specialize the model on a new similar task [45], [81], [82].

There are two main techniques for Transfer Learning:

- *Fine Tuning:* the approach of "fine-tune" the deeper layers of the pre-trained network on the new dataset is typically much faster and easier than training the model from scratch. Although it requires the least amount of data and computational resources [83], the new dataset must be large enough and similar to the pre-trained one.
- *Feature Extraction:* a more specialized method in which data of the new dataset are passed only once through the pre-trained network and then features are extracted from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

one of the pools of the network. These features are then used to train a Machine Learning model such as Support Vector Machine etc. This technique is the most suitable for small datasets.

In this work, the Transfer Learning approach with Feature Extraction was used because of limited dimensions of dataset. In particular, we chose the pre-trained VGGish Convolutional Neural Network (CNN) [84], [85], developed by Google and inspired by the famous VGG networks used for image classification. Its structure consists of a series of convolution and activation layers, optionally followed by a max pooling layer. The VGGish CNN contains 17 layers in total and it is designed for audio classification tasks. Originally, it was employed to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels. In our method, signals were first segmented starting from the first "PA-TA" peak and considering windows with a length of 1 second and 50% overlapped. Then, they were preprocessed to obtain the format required for the network. In particular, they were resampled to 16 kHz, then a one-sided short time Fourier transform was computed, only the magnitude of the complex spectral values was considered discarding the phase. Finally, the Mel spectrogram was calculated and it was converted to a log scale. Overlapped segments of 96 spectra were given in input to the network. Activations of the pooling layer "pool 4" were extracted as features to train machine learning models. We selected "pool 4" since it was the most discriminative pool layer of the pre-trained model of VGGish Convolutional Neural Network [84], [85]. The choice of the pool depends on the similarity between the dataset of the pre-trained model and the dataset of the new application. Since the deeper layers extract higher level features while earlier levels extract lower level ones, the correct depth is as deeper as more similar the datasets are [45], [81], [82]. The structure of VGGish CNN is reported in detail in Table II. The flowchart of features extracted by the VGGish from each data frame of one subject is reported in Fig. 7. We extracted 12288 features from layer "pool 4". After the feature extraction step, we selected the best combination of 1444 deep features using the same approach described in the previous section D for ML. Two variables $PATA_{freq}$ and Age were added also for their high predictive power.

F. CLASSIFICATION

The classification task was conducted processing audio signals as input of a hierarchical model which discerns healthy subjects, low severity patients and high severity patients using Speech Disturbance score of SARA Scale as clinical output. As shown in Fig. 1, we defined binary labels for each level: the first layer to discriminate subjects with Ataxia vs healthy and the second layer trained only on patients to recognize speech disturbance severity (Low [0-1] vs High [2-3]). Speech Disturbance item is one of the eight items that compose SARA scale. It has a score between 0 (normal) - 6 (anarthria) assigned hearing words intelligibility [14].

In our dataset, since the enrolled subjects do not cover the full range of the score, we decided to label it considering the maximum observed value of the 3 to obtain a balanced dataset. Detailed information about the dataset is summarized in Table III. The classification step was performed by testing four of the most conventional classifiers: Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes (NB) and Decision Tree and adopting the majority voting ensemble technique [86]. In our approach, we used two binary levels of classifiers. The first level discriminates healthy vs patients and the second level assesses the speech disturbance severity (Low vs High). We tested the best combination of features extracted with machine learning and deep learning approaches for HMLM, and performed a comparison with a flat classification approach (a parallel multi-classifier with three classes: Healthy vs Low severity vs High severity). Cross-validation techniques such as 5-fold, 10-fold and leave-one-out were applied to check overfitting and to avoid data selection bias. Finally, majority voting ensemble technique was used to aggregate the outputs of the single audio frames into related subjects.

G. PERFORMANCE METRICS

Classification performances were assessed using Accuracy, Precision, Recall and F1-Score [87]. These metrics are summarized in Table IV. For HMLM, the employed definitions of Precision, Recall and F1-Score discriminate and weight differently each type of misclassification error, taking into account the output of each level instead of just the final one [87], [88]. As it regards accuracy, we reported the result for each level and the overall one. Given that cross-validation was carried out, we have computed correctly and incorrectly predictions of each class for each fold and we have summed them up at the end of all iterations before calculating the performance measures Fig. 7.

III. RESULTS

Results about HMLM are shown in Table V. We report the performances of machine learning with canonical features, deep learning features and their combination respectively for level 1 (healthy vs patients) and level 2 (low vs high severity). All the models were tested with ensemble majority voting and three different cross-validation techniques (5-fold, 10-fold and leave-one-out). No significant differences were found between the performance metrics of the three cross-validation methods. The combination of machine learning (level 1) and deep learning (level 2) approaches achieved optimal results in discriminating patients with ataxia from healthy individuals with a mean accuracy of approximately 90%, and in identifying ataxia speech disorders severity with an accuracy of about 80%. For level 1 and for level 2 with 5-fold cross validation (see other details of 10-fold & Leave-one-out in Table V) we achieved a precision of 93.44% and 78.55%, a recall of 98.28% and 79.50% and a f1-score of 95.80% and 79.02% respectively. While the overall precision, recall and f1-score achieved by the model is 84.67%, overall accuracy obtained is 76.32%. Detailed information about collected dataset and classification output is reported in Table VI and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

Fig. 8 with confusion matrix of leave-one-out. We observed that healthy subjects were never classified as patients with high severity, although they were sometimes confused with the low severity patients. Since many of these patients had a speech disturbance score of 0 as healthy subjects and so the two classes were partially overlapped. For the same reason, the network made few mistakes distinguishing patients with low and high severity. Table VII reports results achieved with flat multi-class approach. In this case the use of a unique level with three classes reached a maximum overall accuracy of about 65% with the use of deep learning approach.

IV. DISCUSSION

The aim of this study was to exploit artificial intelligence methods, such as machine learning and the most recent deep learning approaches, to explore and identify new useful strategies from speech analysis and develop innovative reliable and accurate tools for supporting clinical practice in the field of ataxia assessment and treatment. We have explored the possibility of training some automatic predictive models able to identify from audio recordings, the presence of ataxic syndromes and to classify their severity. As far as our knowledge goes, it is the first time in which a HMLM has been applied for the assessment of ataxic disorders with a particular focus on the standardized speech-based "PA-TA" test. The hierarchical approach was investigated in comparison with the flat multi-class approach. The "PA-TA" signal has been pre-processed and segmented and the HMLM has been implemented and tested using two binary classifiers in cascade and adopting the ensemble majority voting technique. We investigated the performances of each level combining conventional and deep learning models. Three combinations of models (machine learning, deep learning, and machine learning + deep learning) were created by using three cross validation approaches as shown in Table V. In Table VII we reported performances of the flat multi-class approach. Results of HMLM showed that the conventional features at the first level work better to classify healthy vs patients, while at the second level the transfer learning features-based method was more suitable to assess the severity of dysarthria. Moreover, the performed experiments demonstrate that the HMLM outperforms the conventional flat classification approach by exhibiting a higher overall accuracy (76.32% vs 65.58%, 69.74% vs 65.89% and 71.05% vs 65.79%) for 5-fold, 10-fold and leave-one-out respectively. Furthermore, the similarity among the three different techniques of cross-validation, speaks about the robustness of our approach. The employed dataset was affected by some limitations such as a relatively small number of available subjects due to the rarity of ataxic syndromes and the lack of variability of speech disturbance score having lower range of severity [0-6] as shown in Table VI. In this scenario, we observed that the HMLM overcomes these aspects performing much better than the widely adopted flat multi-class approach. Despite these limitations, it's also important to say that the collected structured dataset is the first and the biggest released till date, and there are a few works in this field [33]. As evidenced from results of cross validation matrix (Fig. 8), most of the errors were resulted because sometimes the same clinical scores were used for different classes such as healthy and low severity or low and high severity of dysarthria (see also Table VI). This aspect emphasizes how tricky it is for the clinicians to discriminate in scoring the subtle changes of speech dysarthria. These issues highlight the need for larger training datasets for AI-based automatic score annotation. Extensive enrollment of patients will increase statistical variability of severity and the possibility to identify more homogeneous etiopathogenic classes.

V. CONCLUSIONS

This study provided initial evidence on the reliability of digital biomarkers based on speech assessment in the field of ataxia. Specifically, we demonstrated that analysis of "PA-TA" test could provide several variables that are able to accurately classify subjects depending on their conditions (patient or control). From a clinical perspective, these findings have several substantial implications. We introduced a panel of novel objective parameters for clinical evaluation in both observational and interventional contexts, which might turn to be useful as an outcome of measures. Then, the source of the biomarkers (namely, the voice and speech) is such that it may cover patients with ataxia at every disease stage, from early-subclinical to the very advanced, overtaking some limitations of the current assessment systems and being particularly suitable for experimental trials. Finally, such biomarkers will be well fit with the need of implementing telemedicine [89], since voice recording is now possible at distance, by commercial devices, allowing remote monitoring. These results encourage the spread of artificial intelligence in meeting the need of quantitative assessment of disturbances in children with ataxia. An objective evaluation, of what can be clinically relevant in the disease, will contribute to obtaining reliable results also in clinical trials. The association between home-based treatment and devices for the remote monitoring of patients could play a crucial role, in particular, if we think at the efficacy [90], decreasing costs and stress for both patients and their families. Indeed, the HMLM model trained and described in this work could be a powerful tool of telemedicine to be exploited for initial screening and for monitoring in the field of ataxic syndromes, since it requires only an audio recording to assess the conditions of the subject.

REFERENCES

[1] R. Brandsma, T. F. Lawerman, M. J. Kuiper, R. J. Lunsing, H. Burger, and D. A. Sival, "Reliability and discriminant validity of ataxia rating scales in early onset ataxia," *Dev Med Child Neurol*, vol. 59, no. 4, pp. 427–432, Apr. 2017, doi: 10.1111/dmcn.13291.

[2] D. R. Lynch, A. McCormick, K. Schadt, and E. Kichula, "Pediatric Ataxia: Focus on Chronic Disorders," *Seminars in Pediatric Neurology*, vol. 25, pp. 54–64, Apr. 2018, doi: 10.1016/j.spen.2018.01.001.

[3] E. Bertini, G. Zanni, and E. Boltshauser, "Nonprogressive congenital ataxias," *Handb Clin Neurol*, vol. 155, pp. 91–103, 2018, doi: 10.1016/B978-0-444-64189-2.00006-8.

[4] M. Synofzik and A. H. Németh, "Recessive ataxias," *Handb Clin Neurol*, vol. 155, pp. 73–89, 2018, doi: 10.1016/B978-0-444-64189-2.00005-6.

[5] P. Pavone *et al.*, "Ataxia in children: early recognition and clinical evaluation," *Italian Journal of Pediatrics*, vol. 43, no. 1, p. 6, Jan. 2017, doi: 10.1186/s13052-016-0325-9.

[6] T. Schirinzi, A. Sancesario, E. Bertini, E. Castelli, and G. Vasco, "Speech and Language Disorders in Friedreich Ataxia: Highlights on Phenomenology, Assessment, and Therapy," *Cerebellum*, vol. 19, no. 1, pp. 126–130, Feb. 2020, doi: 10.1007/s12311-019-01084-8.

[7] T. Schirinzi *et al.*, "One-year outcome of coenzyme Q10 supplementation in ADCK3 ataxia (ARCA2)," *cerebellum ataxias*, vol. 6, no. 1, p. 15, Dec. 2019, doi: 10.1186/s40673-019-0109-2.

[8] K. E. Musselman *et al.*, "Prevalence of ataxia in children," *Neurology*, vol. 82, no. 1, pp. 80–89, Jan. 2014, doi: 10.1212/01.wnl.0000438224.25600.6c.

[9] J. López-Bastida, L. Perestelo-Pérez, F. Montón-álvarez, and P. Serrano-Aguilar, "Social economic costs and health-related quality of life in patients with degenerative cerebellar ataxia in Spain," *Movement Disorders*, vol. 23, no. 2, pp. 212–217, 2008, doi: https://doi.org/10.1002/mds.21798.

[10] T. Schirinzi *et al.*, "Childhood Rapid-Onset Ataxia: Expanding the Phenotypic Spectrum of ATP1A3 Mutations," *Cerebellum*, vol. 17, no. 4, pp. 489–493, Aug. 2018, doi: 10.1007/s12311-018-0920-y.

[11] T. Schirinzi *et al.*, "Non-invasive Focal Mechanical Vibrations Delivered by Wearable Devices: An Open-Label Pilot Study in Childhood Ataxia," *Front Neurol*, vol. 9, p. 849, 2018, doi: 10.3389/fneur.2018.00849.

[12] M. B. Delatycki and S. I. Bidichandani, "Friedreich ataxia-pathogenesis and implications for therapies," *Neurobiol Dis*, vol. 132, p. 104606, Dec. 2019, doi: 10.1016/j.nbd.2019.104606.

[13] A. Benussi, A. Pascual-Leone, and B. Borroni, "Non-Invasive Cerebellar Stimulation in Neurodegenerative Ataxia: A Literature Review," *Int J Mol Sci*, vol. 21, no. 6, Mar. 2020, doi: 10.3390/ijms21061948.

[14] T. Schmitz-Hübsch *et al.*, "Scale for the assessment and rating of ataxia: development of a new clinical scale," *Neurology*, vol. 66, no. 11, pp. 1717–1720, Jun. 2006, doi: 10.1212/01.wnl.0000219042.60538.92.

[15] J. Folker, B. Murdoch, L. Cahill, M. Delatycki, L. Corben, and A. Vogel, "Dysarthria in Friedreich's ataxia: a perceptual analysis," *Folia Phoniatr Logop*, vol. 62, no. 3, pp. 97–103, 2010, doi: 10.1159/000287207.

[16] M. Germanotta *et al.*, "Robotic and clinical evaluation of upper limb motor performance in patients with Friedreich's Ataxia: an observational study," *Journal of NeuroEngineering and Rehabilitation*, vol. 12, no. 1, p. 41, Apr. 2015, doi: 10.1186/s12984-015-0032-6.

[17] T. F. Lawerman, R. Brandsma, H. Burger, J. G. M. Burgerhof, D. A. Sival, and the Childhood Ataxia and Cerebellar Group of the European Pediatric Neurology Society, "Age-related reference values for the pediatric Scale for Assessment and Rating of Ataxia: a multicentre study," *Dev Med Child Neurol*, vol. 59, no. 10, pp. 1077–1082, Oct. 2017, doi: 10.1111/dmcn.13507.

[18] R. Brandsma *et al.*, "Ataxia rating scales are age-dependent in healthy children," *Developmental Medicine & Child Neurology*, vol. 56, no. 6, pp. 556–563, 2014, doi: https://doi.org/10.1111/dmcn.12369.

[19] S. Summa *et al.*, "Validation of low-cost system for gait assessment in children with ataxia," *Comput Methods Programs Biomed*, vol. 196, p. 105705, Nov. 2020, doi: 10.1016/j.cmpb.2020.105705.

[20] S. Summa *et al.*, "Spatio-temporal parameters of ataxia gait dataset obtained with the Kinect," *Data Brief*, vol. 32, p. 106307, Oct. 2020, doi: 10.1016/j.dib.2020.106307.

[21] G. Arcuria, C. Marcotulli, C. Galasso, F. Pierelli, and C. Casali, "15-White Dots APP-Coo-Test: a reliable touch-screen application for assessing upper limb movement impairment in patients with cerebellar ataxias," *J Neurol*, vol. 266, no. 7, pp. 1611–1622, Jul. 2019, doi: 10.1007/s00415-019-09299-9.

[22] G. Arcuria *et al.*, "Developing a smartphone application, triaxial accelerometer-based, to quantify static and dynamic balance deficits in patients with cerebellar ataxias," *J Neurol*, vol. 267, no. 3, pp. 625–639, Mar. 2020, doi: 10.1007/s00415-019-09570-z.

[23] H. Tran, P. N. Pathirana, M. Horne, L. Power, and D. Szmulewicz, "Automated Finger Chase (ballistic tracking) in the Assessment of Cerebellar Ataxia," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Oct. 2018, pp. 3521–3524. doi: 10.1109/EMBC.2018.8513036.

[24] H. Tran, K. D. Nguyen, P. N. Pathirana, M. K. Horne, L. Power, and D. J. Szmulewicz, "A comprehensive scheme for the objective upper body assessments of subjects with cerebellar ataxia," *J Neuroeng Rehabil*, vol. 17, Dec. 2020, doi: 10.1186/s12984-020-00790-3.

[25] H. Tran, K. D. Nguyen, P. N. Pathirana, M. Horne, L. Power, and D. J. Szmulewicz, "Multimodal Data Acquisition for the Assessment of Cerebellar Ataxia via Ballistic Tracking," *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2020, pp. 859–862, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176379.

[26] E. Lacorte, G. Bellomo, S. Nuovo, M. Corbo, N. Vanacore, and P. Piscopo, "The Use of New Mobile and Gaming Technologies for the Assessment and Rehabilitation of People with Ataxia: a Systematic Review and Meta-analysis," *Cerebellum*, Nov. 2020, doi: 10.1007/s12311-020-01210-x.

[27] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly, and P. J. Snyder, "Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment," *Journal of Neurolinguistics*, vol. 17, no. 6, pp. 439–453, Nov. 2004, doi: 10.1016/j.jneuroling.2004.06.001.

[28] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biol Psychiatry*, vol. 72, no. 7, pp. 580–587, Oct. 2012, doi: 10.1016/j.biopsych.2012.03.015.

[29] A. P. Vogel, C. Shirbin, A. J. Churchyard, and J. C. Stout, "Speech acoustic markers of early stage and prodromal Huntington's disease: a marker of disease onset?," *Neuropsychologia*, vol. 50, no. 14, pp. 3273–3278, Dec. 2012, doi: 10.1016/j.neuropsychologia.2012.09.011.

[30] A. P. Vogel, J. Fletcher, P. J. Snyder, A. Fredrickson, and P. Maruff, "Reliability, stability, and sensitivity to change and impairment in acoustic measures of timing and frequency," *J Voice*, vol. 25, no. 2, pp. 137–149, Mar. 2011, doi: 10.1016/j.jvoice.2009.09.003.

[31] A. P. Vogel and P. Maruff, "Monitoring change requires a rethink of assessment practices in voice and speech," *Logoped Phoniatr Vocol*, vol. 39, no. 2, pp. 56–61, Jul. 2014, doi: 10.3109/14015439.2013.775332.

[32] A. P. Vogel *et al.*, "Speech and swallowing abnormalities in adults with POLG associated ataxia (POLG-A)," *Mitochondrion*, vol. 37, pp. 1–7, Nov. 2017, doi: 10.1016/j.mito.2017.06.002.

[33] N. P. Narendra and P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Computer Speech & Language*, vol. 65, p. 101117, Jan. 2021, doi: 10.1016/j.csl.2020.101117.

[34] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, 2016, doi: 10.1016/j.bbe.2015.11.004.

[35] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 4605–4608. doi: 10.1109/ICASSP.2009.4960656.

[36] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Comput Speech Lang*, vol. 29, no. 1, pp. 132–144, Jan. 2015, doi: 10.1016/j.csl.2014.02.001.

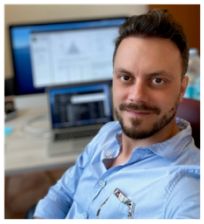
[37] M. S. Paja and T. H. Falk, "Automated Dysarthria Severity Classification for Improved Objective Intelligibility Assessment of Spastic Dysarthric Speech," p. 4.

[38] S. A. Selouani, H. Dahmani, R. Amami, and H. Hamam, "Using speech rhythm knowledge to improve dysarthric speech recognition," *International Journal of Speech Technology*, vol. 15, Mar. 2012, doi: 10.1007/s10772-011-9104-6.

[39] F. Rudzicz *et al.*, "ORIGINAL PAPER The TORGO database of acoustic and articulatory speech from speakers with dysarthria." 2011.
[40] J. B. Polikoff and H. T. Bunnell, "THE NEMOURS DATABASE OF DYSARTHRIC SPEECH: A PERCEPTUAL ANALYSIS," p. 4.
[41] S. Summa *et al.*, "Development of SaraHome: A novel, well-accepted, technology-based assessment tool for patients with ataxia," *Comput Methods Programs Biomed*, vol. 188, p. 105257, May 2020, doi: 10.1016/j.cmpb.2019.105257.
[42] D. R. Lynch *et al.*, "Measuring Friedreich ataxia: Complementary features of examination and performance measures," *Neurology*, vol. 66, no. 11, pp. 1711–1716, Jun. 2006, doi: 10.1212/01.wnl.0000218155.46739.90.
[43] S. H. Subramony *et al.*, "Measuring Friedreich ataxia: Interrater reliability of a neurologic rating scale," *Neurology*, vol. 64, no. 7, pp. 1261–1262, Apr. 2005, doi: 10.1212/01.WNL.0000156802.15466.79.
[44] M. Asad Ullah and S. Nisar, "A Silence Removal and Endpoint Detection Approach for Speech Processing," Sep. 2016.
[45] F. Ponzio, E. Macii, E. Ficarra, and S. Di Cataldo, "Colorectal Cancer Classification using Deep Convolutional Networks - An Experimental Study:," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, Funchal, Madeira, Portugal, 2018, pp. 58–66. doi: 10.5220/0006643100580066.
[46] F. Alías, J. C. Socoró, and X. Sevillano, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds," 2016, doi: 10.3390/APP6050143.
[47] S. Pincus, "Approximate Entropy as a Measure of System Complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, pp. 2297–301, Apr. 1991, doi: 10.1073/pnas.88.6.2297.
[48] R. Metzger, J. Doherty, and D. Jenkins, "Using Approximate Entropy as a speech quality measure for a speaker recognition system," Mar. 2016, pp. 292–297. doi: 10.1109/CISS.2016.7460517.
[49] M. Banbrook and S. McLaughlin, "Is speech chaotic?: invariant geometrical measures for speech data," in *IEE Colloquium on Exploiting Chaos in Signal Processing*, Jun. 1994, p. 8/1-810.
[50] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Springer Netherlands, 1990, pp. 241–261. doi: 10.1007/978-94-009-2037-8_10.
[51] F. González, A. Guillamón, J. C. Alcaraz, and M. C. Alcaraz, "Detection of chaotic behaviour in speech signals using the largest Lyapunov exponent," Feb. 2002, vol. 1, pp. 317–320 vol.1. doi: 10.1109/ICDSP.2002.1027895.
[52] V. Pitsikalis, I. Kokkinos, and P. Maragos, "Nonlinear analysis of speech signals: Generalized dimensions and Lyapunov exponents," Jan. 2003.
[53] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–17, Jan. 1999, doi: 10.1109/89.736326.
[54] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral Entropy Based Feature for Robust ASR," Jun. 2004, p. I–193. doi: 10.1109/ICASSP.2004.1325955.
[55] "Peeters_2003_cuidadoaudiofeatures.pdf." Accessed: Mar. 24, 2021. [Online]. Available: http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
[56] M. Darji, "Audio Signal Processing: A Review of Audio Signal Classification Features," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, pp. 227–230, May 2017.
[57] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012, doi: 10.1109/TMM.2012.2199972.
[58] "Spectral centroid for audio signals and auditory spectrograms - MATLAB spectralCentroid - MathWorks Italia." https://it.mathworks.com/help/audio/ref/spectralcentroid.html (accessed Oct. 18, 2021).
[59] "Spectral spread for audio signals and auditory spectrograms - MATLAB spectralSpread - MathWorks Italia." https://it.mathworks.com/help/audio/ref/spectralspread.html (accessed Oct. 18, 2021).
[60] "Spectral skewness for audio signals and auditory spectrograms - MATLAB spectralSkewness - MathWorks Italia." https://it.mathworks.com/help/audio/ref/spectralskewness.html (accessed Oct. 18, 2021).
[61] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 282–307, Feb. 2006, doi: 10.1016/j.ymssp.2004.09.001.
[62] J. Antoni and R. B. Randall, "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 308–331, Feb. 2006, doi: 10.1016/j.ymssp.2004.09.002.
[63] "Spectral kurtosis from signal or spectrogram - MATLAB pkurtosis - MathWorks Italia." https://it.mathworks.com/help/signal/ref/pkurtosis.html (accessed Oct. 18, 2021).
[64] "An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics | Wiley," *Wiley.com*. https://www.wiley.com/en-us/An+Introduction+to+Audio+Content+Analysis%3A+Applications+in+Signal+Processing+and+Music+Informatics-p-9781118266823 (accessed Oct. 18, 2021).
[65] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, vol. 2, pp. 1331–1334. doi: 10.1109/ICASSP.1997.596192.
[66] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas Commun.*, 1988, doi: 10.1109/49.608.
[67] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, vol. 1, p. I–193. doi: 10.1109/ICASSP.2004.1325955.
[68] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," *The Journal of the Acoustical Society of America*, vol. 52, no. 6B, pp. 1687–1697, Dec. 1972, doi: 10.1121/1.1913303.
[69] S. Gonzalez and M. Brookes, "A Pitch Estimation Filter Robust to High Levels of Noise (PEFAC)," p. 5.
[70] D. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, vol. 83, pp. 257–64, Feb. 1988, doi: 10.1121/1.396427.
[71] T. Drugman and A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," Jan. 2011, pp. 1973–1976.
[72] "MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval | Wiley," *Wiley.com*. https://www.wiley.com/en-am/MPEG+7+Audio+and+Beyond%3A+Audio+Content+Indexing+and+Retrieval-p-9780470093344 (accessed Oct. 18, 2021).
[73] "Quadratic Interpolation of Spectral Peaks." https://ccrma.stanford.edu/~jos/sasp/Quadratic_Interpolation_Spectral_Peaks.html (accessed Oct. 18, 2021).
[74] M. Reddy and L. Reddy, "Dimensionality Reduction: An Empirical Study on the Usability of IFE-CF (Independent Feature Elimination- by C-Correlation and F-Correlation) Measures," *International Journal of Computer Science Issues*, vol. 7, Feb. 2010.
[75] I. Guyon and A. Elisseeff, "An Introduction of Variable and Feature Selection," *J. Machine Learning Research Special Issue on Variable and Feature Selection*, vol. 3, pp. 1157–1182, Jan. 2003, doi: 10.1162/153244303322753616.
[76] O. Al-Harbi, "A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine," *arXiv:1902.06242 [cs]*, Feb. 2019, Accessed: Mar. 24, 2021. [Online]. Available: http://arxiv.org/abs/1902.06242
[77] S. An and X. Fan, "Study on Method of Feature Selection in Speech Content Classification," *IJACSA*, vol. 5, no. 4, 2014, doi: 10.14569/IJACSA.2014.050412.
[78] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Nov. 2018, pp. 160–163. doi:

10.1109/ICSESS.2018.8663882.
[79] J. R. Delgado-Contreras, J. P. Garćıa-Vázquez, R. F. Brena, C. E. Galván-Tejada, and J. I. Galván-Tejada, "Feature Selection for Place Classification through Environmental Sounds," *Procedia Computer Science*, vol. 37, pp. 40–47, Jan. 2014, doi: 10.1016/j.procs.2014.08.010.
[80] K. Weiss, T. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, May 2016, doi: 10.1186/s40537-016-0043-6.
[81] "Extract Image Features Using Pretrained Network - MATLAB & Simulink - MathWorks Italia." https://it.mathworks.com/help/deeplearning/ug/extract-image-features-using-pretrained-network.html (accessed Mar. 24, 2021).
[82] "Pretrained Deep Neural Networks - MATLAB & Simulink - MathWorks Italia." https://it.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html (accessed Mar. 24, 2021).
[83] "Rete neurale convoluzionale." https://it.mathworks.com/discovery/convolutional-neural-network-matlab.html (accessed Oct. 18, 2021).
[84] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135. doi: 10.1109/ICASSP.2017.7952132.
[85] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 776–780. doi: 10.1109/ICASSP.2017.7952261.
[86] M. Re and G. Valentini, "Ensemble methods: A review," in *Advances in Machine Learning and Data Mining for Astronomy*, 2012, pp. 563–594.
[87] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[88] S. Kiritchenko, S. Matwin, R. Nock, and A. F. Famili, "Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization," in *Advances in Artificial Intelligence*, vol. 3060, A. Y. Tawfik and S. D. Goodwin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 395–406. doi: 10.1007/11766247_34.
[89] T. Schirinzi, A. Sancesario, E. Castelli, E. Bertini, and G. Vasco, "Friedreich ataxia in COVID-19 time: current impact and future possibilities," *cerebellum ataxias*, vol. 8, no. 1, p. 4, Dec. 2021, doi: 10.1186/s40673-020-00127-9.
[90] A. P. Vogel *et al.*, "Speech treatment improves dysarthria in multisystemic ataxia: a rater-blinded, controlled pilot-study in ARSACS," *J Neurol*, vol. 266, no. 5, pp. 1260–1266, May 2019, doi: 10.1007/s00415-019-09258-4.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access



GENNARO TARTARISCO received a MSc in Biomedical Engineering at the University of Pisa in 2009 and a Ph.D. degree in Automatic, Robotic and Bioengineering from the same university in 2013. Dr. Tartarisco was a Research Fellow at the Institute of Clinical physiology of National Research Council of Italy, in collaboration with the Interdepartmental Research Centre "E. Piaggio" of the Faculty of Engineering of University of Pisa from 2013 to 2015. Since 2016 He is a PhD Researcher at the Institute for Biomedical Research and Innovation (IRIB) of National Research Council of Italy, Messina unit. His research approach spans from mobile health technology and computational science in medicine. His interests lie around the recent advances of wearable healthcare systems and telemedicine, coupling with signal processing and data mining techniques such as machine learning and deep learning algorithms. One of the main contributions of his research works is related to the monitoring of health parameters, as physiological and cardiovascular clues, vital body signs and the patients' social and environmental context to understand biological phenomena and assisting medical diagnosis, rehabilitation and early detection of diseases. He has published over 60 works in peer-reviewed international journals.

ROBERTA BRUSCHETTA was born in Messina, Italy in 1994. She received the M.S. degree in Biomedical Engineering cum laude from Politecnico of Turin, Italy in 2019. During her education she gained experience in bioimages and signal processing. She is currently a Research Fellow at the Institute for Biomedical Research and Innovation (IRIB) of National Research Council of Italy, Messina unit. She has contributed to researches concerning the development of automatic systems for early diagnosis of neurodevelopmental disorders and identification of new biomarkers through signal and image processing techniques and data mining approaches such as machine learning and deep learning. Her research interests range from the field of ICT- based medical devices design to the artificial intelligence and the development of medical software and algorithms to support clinical decisions.

SUSANNA SUMMA is a member of the Movement Analysis and Robotic laboratory (MARlab), at the Bambino Gesù Children Hospital under the supervision of Maurizio Petrarca. Previously she was at the Research Unit of Neurophysiology & Neuroengineering of Human Technology Interaction, at the Campus Bio-Medico University. And before at the NeuroEngineering and NeuroRobotics Lab (Neurolab) of University of Genova (Unige), under the supervision of Vittorio Sanguineti and Maura Casadio. Her main research areas are neural control of movements, motor learning in normal and pathologic conditions, neuromotor recovery through robotic technologies and natural interfaces in persons with neurological disorders (stroke, multiple sclerosis, Parkinson's disease, spinal cord injury, ataxia). Currently she is focusing her interests in the development of novel wearable technologies, in standard movement analysis of pediatric patients and in the study of human-machine interaction as tools for the assessment of neuromotor diseases and for neuromotor recovery.

LILIANA RUTA is a child neuropsychiatrist. During her clinical and research training and Ph.D, she has developed an extensive experience on neurodevelopmental conditions with a particular focus on autism, from biology to clinical implications. From 2007 to 2012 she was a visiting Ph.D student and research associate at the Autism Research Centre (Cambridge University, UK), where she consolidated a sound expertise in early behavioural and neurocognitive markers of autism and early diagnosis. In her current research position at the CNR, she is leading a research group developing and carrying out novel experimental paradigms to examine young children with autism and to track their development in all the domains. In the last few years, she started a new research line aimed to explore the efficacy and feasibility of an early parent-mediated intervention based on ESDM, implemented through the use of tech-enabled remote monitoring (telehealth).

MARTINA FAVETTA graduated in Developmental Therapy, at the University of Rome "La Sapienza" in 2014. She received master's degree in Rehabilitation Science, at the University of Rome "La Sapienza" in 2017. She is currently a researcher at the Movement Analysis and Robotics Laboratory (MARlab), Department of intensive Neurorehabilitation and Robotics, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy. Her current research interests including neuroscience, robotics and pediatric neurorehabilitation. She conducts assessments with 3D system of Gait Analysis and clinical assessment in several neuromotor diseases and child disabilities.

MARIO BUSA' received a MSc in Computer Science in 2011 and a Ph.D. degree in Cognitive Science in 2017 at the University of Messina. During his university career he acquired skills in software development and data science techniques. Since 2020 he is a Research Fellow at the Institute for Biomedical Research and Innovation (IRIB) of National Research Council of Italy. He has contributed to the development of health care mobile applications for the screening of hearth diseases and early diagnosis of neurodevelopment disorders through signal processing and robotic systems. His research interests cover different fields such as robotics and e-health mobile applications for data collection and analysis.

ALBERTO ROMANO is a developmental therapist. He obtained a bachelor's degree in Developmental Neuropsychomotor therapy at the University of Pavia, a first level University Master in Applied Behavior Analysis (ABA) at the University of Parma, and a Master's Degree in Rehabilitation Sciences at the University of Rome "Tor Vergata". He is a founder and member of the board of directors of SMART ONLUS (www.centrosmart.it) since 2016, it is a rehabilitation centre for evidence-based pediatric rehabilitation in the field of behavioral sciences. Since 2017 he carryed out research activities as developmental therapist at the Movement Analysis and Robotics Laboratory (MARLab) of the Bambino Gesù Pediatric Hospital, Department of Neuroscience and Neurorehabilitation. He is a consultant as a therapist and researcher for the Italian Rett Syndrome Association (AIRett) since 2013. He carries out weekly rehabilitation activities with children with neurodevelopmental disabilities. His main research interests concern the practice of neurorehabilitation in developmental age, movement analysis and the application of rehabilitation technologies.

ENRICO CASTELLI Director of the Paediatric Neurorehabilitation Unit and of the Unit for Severe Developmental Disabilities of the Department of Neuroscience and Neurorehabilitation of the Bambino Gesù Hospital in Rome. He graduated in medicine in 1983 and specialised in physical medicine and rehabilitation in 1986 and in neurology in 1991. He is the author of more than 70 publications in the field of neurology and rehabilitation, including original articles and book chapters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

FLAVIA MARINO graduated in Clinical Psychology at the University of Messina in 2008 and received a Ph.D. degree in Psychological Sciences from the same university in 2013. Since 2016 she is a PhD Researcher at the Institute for Biomedical Research and Innovation (IRIB) of National Research Council of Italy, Messina unit. Her research activity is focused on human-technology interaction in order to understand how interactive technologies can be used to enhance cognitive and socio-emotional processes in individuals with disabilities. She is also interested in the study of the cognitive and psychological processes that determine the emotional-relational development of individuals with neuro-disorders with the aim of developing innovative methodologies and tools for the study and strengthening of cognitive abilities and mental health. She is currently involved in observational studies and clinical trials concerning patients with Autism, patients with Congestive Heart Failure, Neurovegetative diseases and preventive interventions for psychophysical wellbeing of elder. She has received an honorary fellowship in Psychology.

ANTONIO CERASA is a cognitive and translational neuroscientist, interested in applying basic neuroscience research into clinical applications for the treatment of neurological and psychiatric disorders. His expertise is characterized by a broad background in neurobiology and neurophysiology of neurological and psychiatric disorders, with specific training in neuropsychology. As PI or co-Investigator on some Italian and European-funded grants, he laid the groundwork for the development of several advanced neuroimaging measures of neurological disorders. Author and co-author of more than 160 publications index in MedLine [H-Index: 43], he's received several International awards (from OHBM and ECTRIMS) and a national prize for science dissemination.

TOMMASO SCHIRINZI is neurologist, assistant professor of neurology at the University of Roma Tor Vergata. His area of interest covers movement disorders and neurodegenerative diseases. He is involved in biomarkers research, either as fluid or digital.

MAURIZIO PETRARCA PhD is a clinician that is the head of Movement Analysis and Robotics Laboratory (MARLab). His actions were addressed at the introduction of technologies in the neurological rehabilitation fields since 1985, when he was a pioneering developer of a computerized optoelectronic system for movement analysis in rehabilitative clinics that received a prize in 1991. He successively patented an elastic component for gait rehabilitation in 1995 and a knee and ankle robotic orthosis in 2012. Currently he is developing a system for sensory-motor study and training, in neurological diseases, that integrates Movement Analysis, 6 DoF Stewart Robotic Platform, Immersive Virtual Reality and robotic orthosis.

ENRICO BERTINI is a pediatric neurologist leading the Clinical, Diagnostic and Research Laboratory Unit of Neuromuscolar and Neurodegenerative Diseases, and the Laboratory of Molecolar Medicine of the Bambino Gesù Children's Hospital, in Rome. He is Contarct Professor of Pediatric Neurology, and, has consolidated and worldwide known expertise in neuromuscular disorders.

GESSICA VASCO is a pediatric neurologist at Department of Neuroscience and Neurorehabilitation of OPBG. She is the main collaborator of the Bioengineer center of OPBG called MARlab, dedicated to measure motor and coordination performances. She has a very good experience in observational and treatment clinical trials in several pediatric neurological disease, neuromuscular and neurodegenerative diseases.

GIOVANNI PIOGGIA Electronic Engineer with specialization in Bioengineering, PhD in Robotics, is a Senior Researcher of the National Research Council (CNR), Institute for Biomedical Research and Innovation (IRIB), Manager of the Messina secondary unit. His research activities are focused on health challenges. The domains of his research are biomedical sensing, early diagnosis and treatment, digital therapeutics, social robotics and computational science. Health challenges are mainly related to neurodevelopmental disorders, with particular attention to autism spectrum disorders, and disabilities. The exploration of digital therapeutics and social robots is dedicated to social-emotional processing, rehabilitation and evidence-based software intervention in order to help people with communication difficulties to learn, identify and use emotional information, and to adapt to social context.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

FEATURE TIPOLOGY	FEATURE NAME
Time-domain features	• PATA frequency • Approximate Entropy
Chaotic features	• Lyapunov Exponent
Frequency-domain features	• 13 mfcc coefficients • 13 gtcc coefficients • Spectral Centroid • Spectral Crest • Spectral Decrease • Spectral Entropy • Spectral Flatness • Spectral Kurtosis • Spectral RolloffPoint • Spectral Skewness • Spectral Slope • Spectral Spread • Pitch • Harmonic Ratio
Other features	• Age

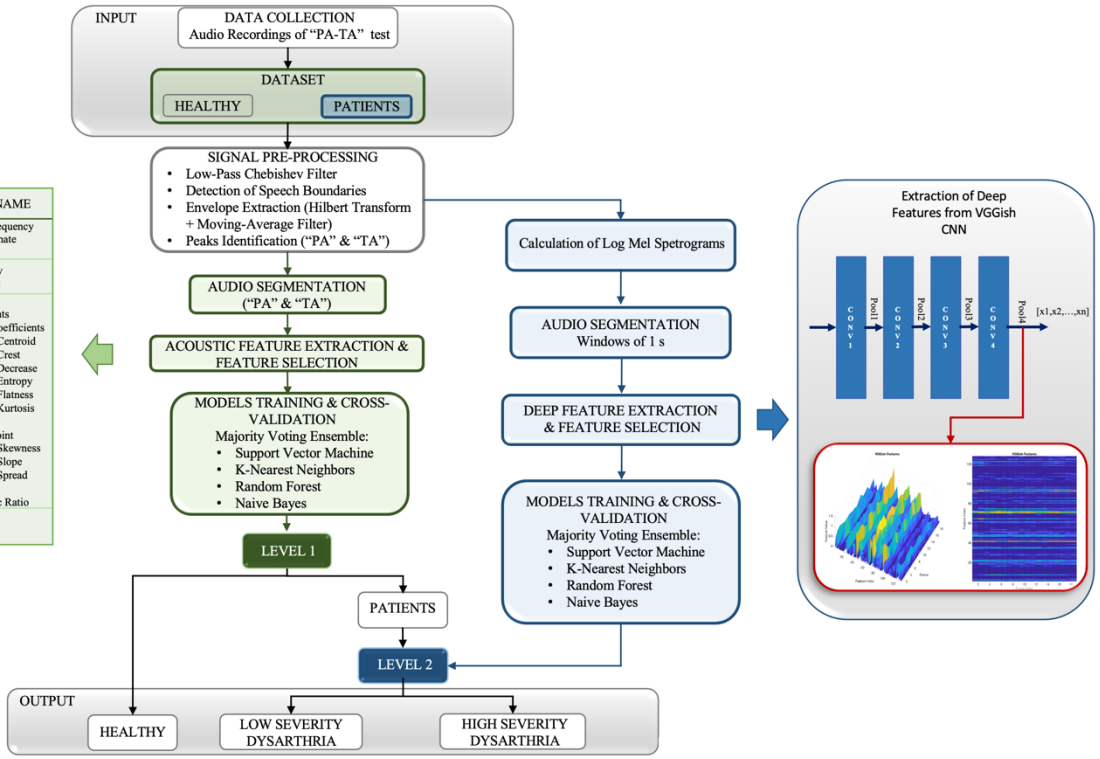


FIGURE 1. Flowchart of overall architecture

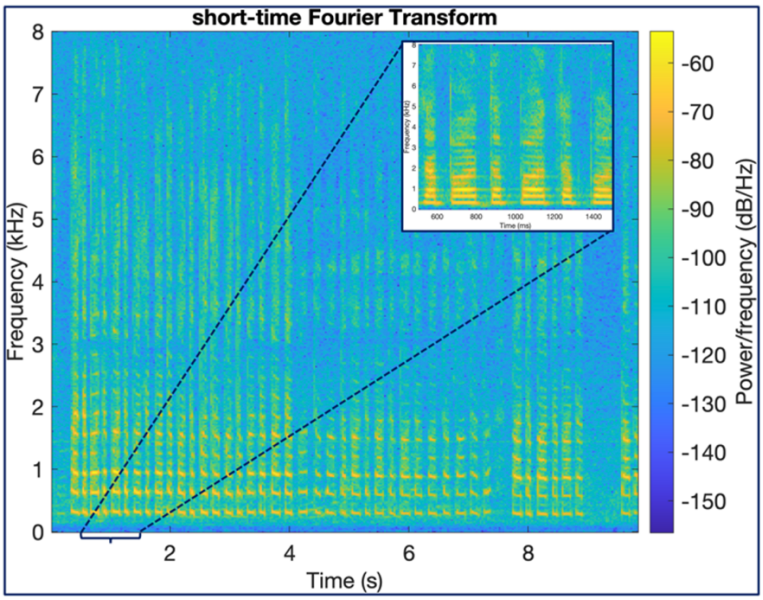


FIGURE 2. Signal short-time Fourier Transform with a detailed view between 500 ms and 1500 ms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

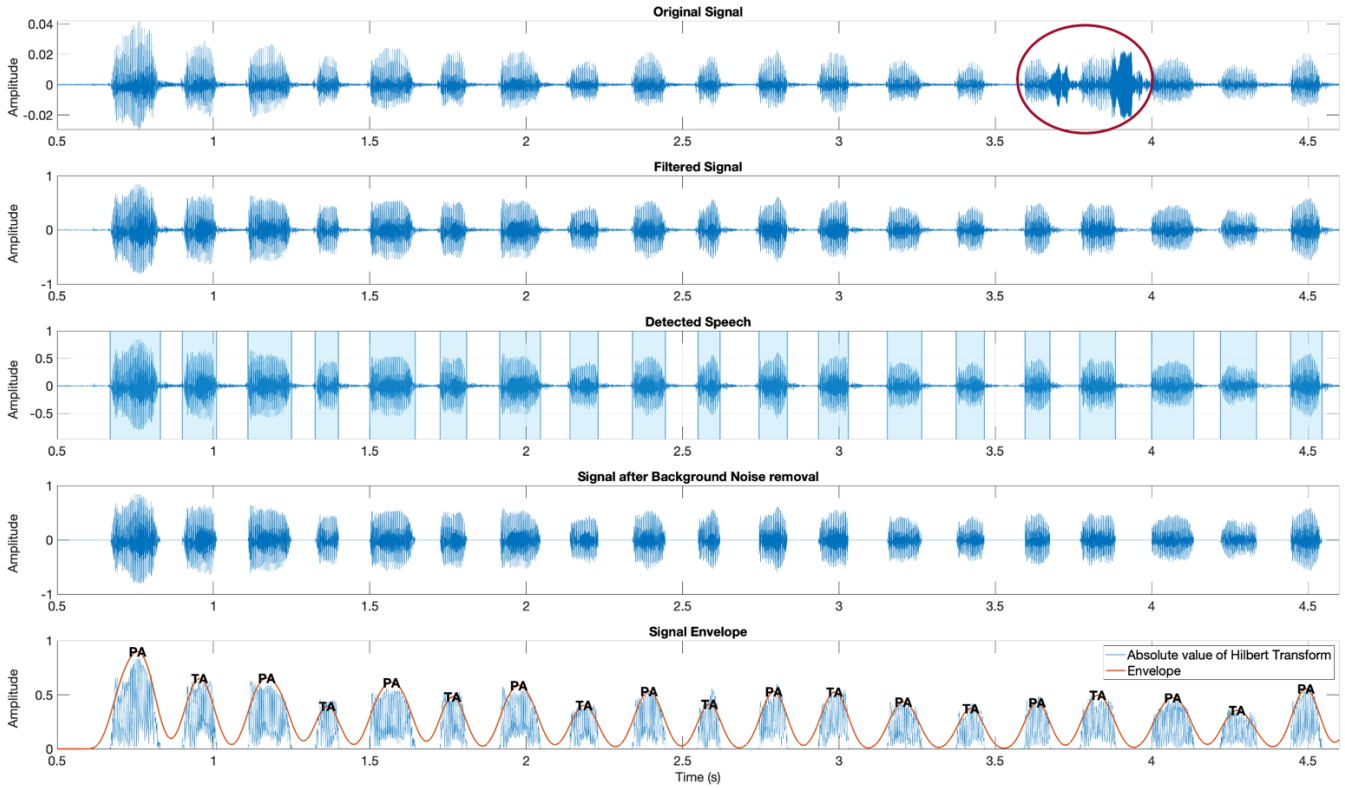


FIGURE 3. Main Steps of signal pre-processing from noise removal to "PA" & "TA" peaks identification. The red circle on original signal highlights an example of background noise (squeak) that is removed through the low-pass filter. Later, the detect speech method based on threshold short-term energy and spectral spread [44] is employed to remove the remaining background noise. The cleaned-up signal is used to obtain the envelope and consequently to identify peaks.

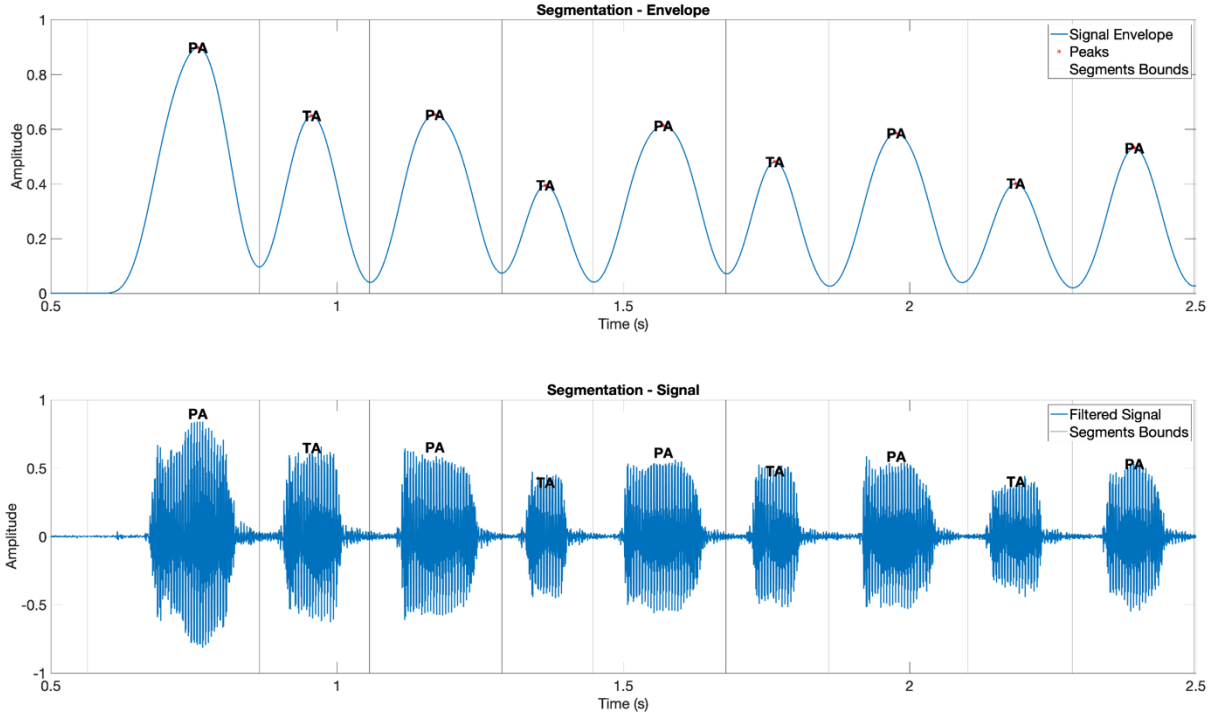


FIGURE 4. Audio segmentation. Each signal was segmented using "PA" & "TA" peaks identified on the envelope as reference points and considering, for each PA-TA cycle, only the samples between the closest preceding and consecutive minima. Each peak corresponds to a syllable.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

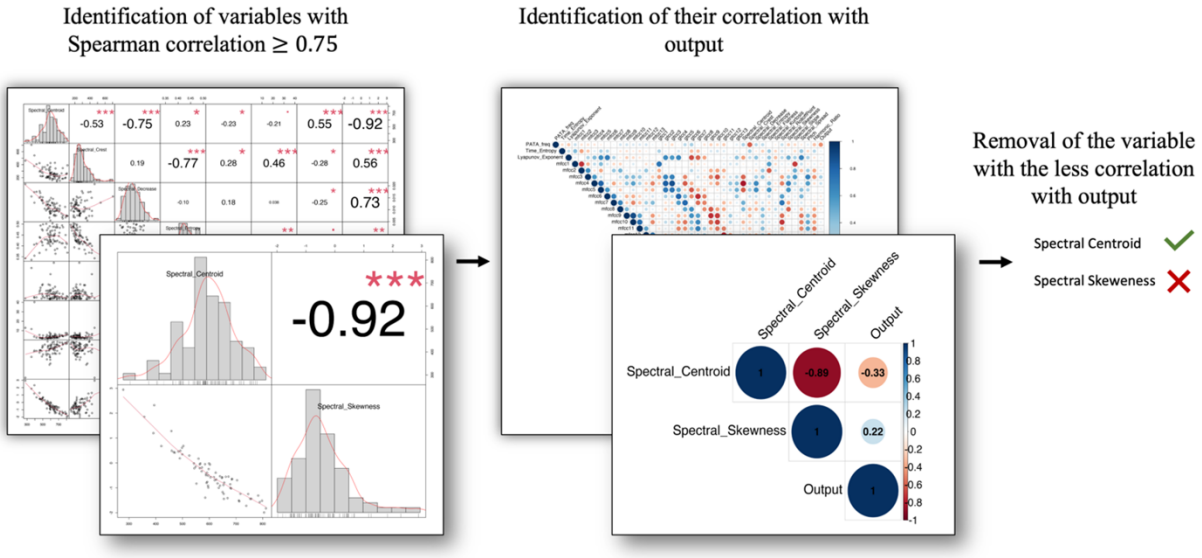


FIGURE 5. **First Feature Selection step based on remotion of variables with Spearman correlation ≥ 75% and lowest correlation with the output. As example we report the remotion of Spectral Skeweness with correlation of 89% with Spectral Centroid and 22% with output classes.**

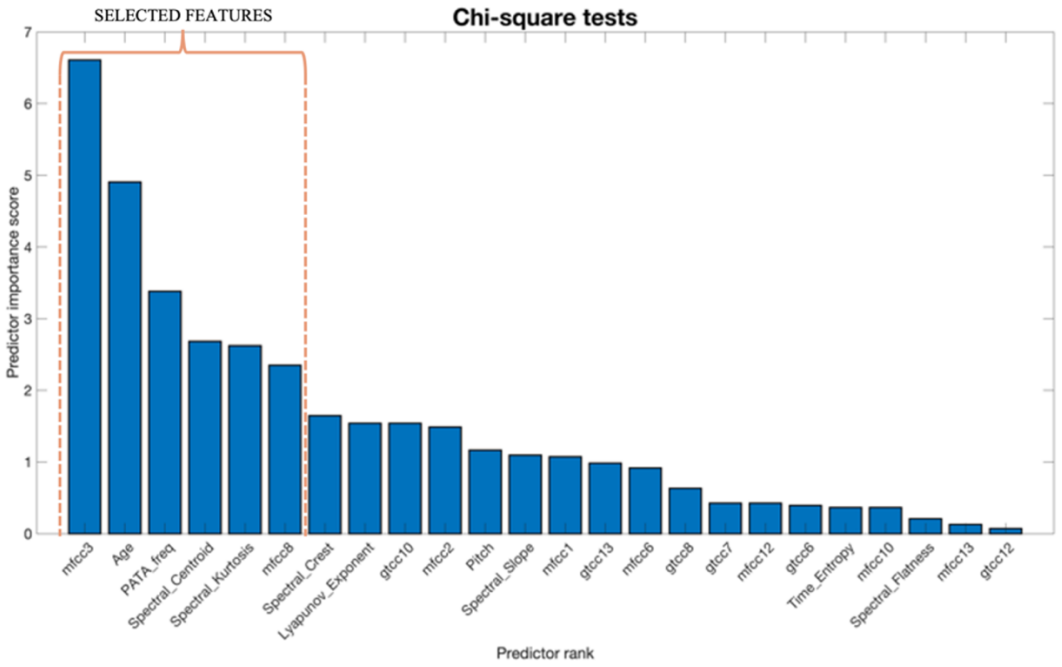


FIGURE 6. **Second Feature Selection step based on Chi-square test. Features are ranked and the break-point is chosen as the highest difference between consecutive scores with the constraint of at least 3 features.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

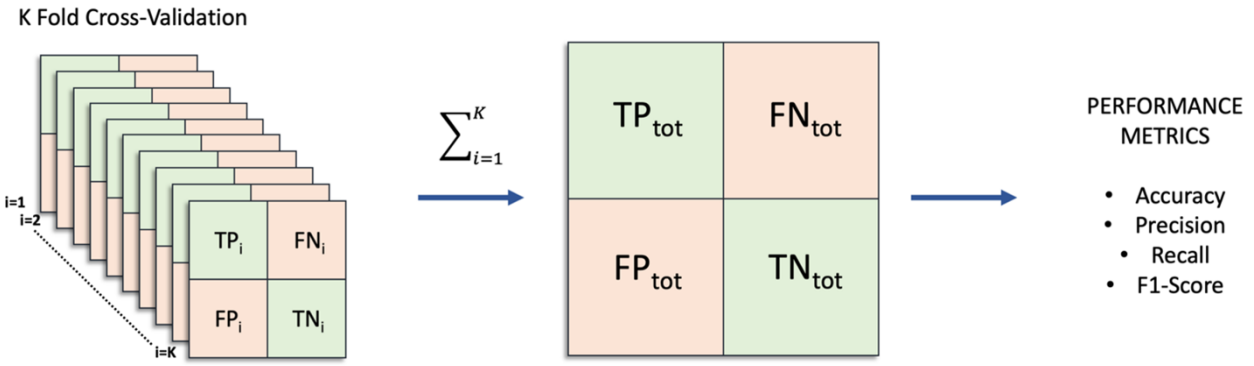


FIGURE 7. **K Folds results aggregation. For each typology of cross-validation, we created a unique confusion matrix and then we computed the performance measures.**

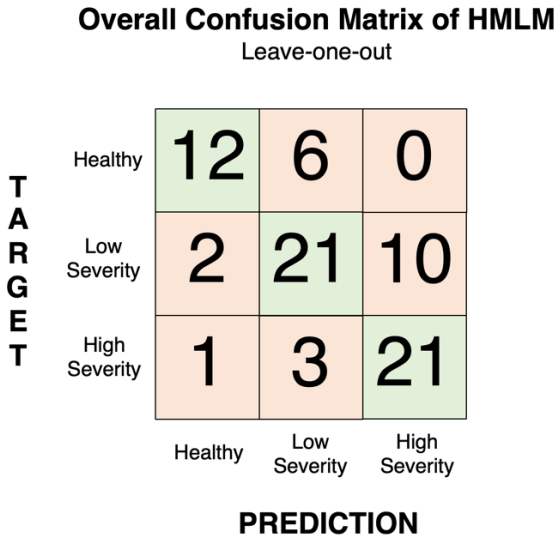


FIGURE 8. **Final confusion matrix of the Hierarchical model using leave-one out validation.**

TABLE I
MOVING-AVERAGE FILTER PARAMETERS

Approximate Entropy Threshold		Numerator Coefficients of the Rational Transfer Function (b)	Denominator Coefficients of the Rational Transfer Function (a)
Entropy < 0.8		$\overrightarrow{(c)}_n$ where c=1/1500, n=1500	0.7
Entropy ≥ 0.8	Filter 1	$\overrightarrow{(c)}_n$ where c=1/1000, n=1000	0.7
	Filter 2	$\overrightarrow{(c)}_n$ where c=1/500, n=500	1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

TABLE II
VGGISH NETWORK STRUCTURE

	Layer Name	Layer type	Layers details
1	'InputBatch'	Image Input	96x64x1 images
2	'conv1'	Convolution	64 3x3x1 convolutions with stride [1 1] and padding "same"
3	'relu'	ReLU	ReLU
4	'pool1'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
5	'conv2'	Convolution	128 3x3x64 convolutions with stride [1 1] and padding "same"
6	'relu2'	ReLU	ReLU
7	'pool2'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
8	'conv3_1'	Convolution	256 3x3x128 convolutions with stride [1 1] and padding "same"
9	'relu3_1'	ReLU	ReLU
10	'conv3_2'	Convolution	256 3x3x256 convolutions with stride [1 1] and padding "same"
11	'relu3_2'	ReLU	ReLU
12	'pool3'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
13	'conv4_1'	Convolution	512 3x3x256 convolutions with stride [1 1] and padding "same"
14	'relu4_1'	ReLU	ReLU
15	'conv4_2'	Convolution	512 3x3x512 convolutions with stride [1 1] and padding "same"
16	'relu4_2'	ReLU	ReLU
17	'pool4'	Max Pooling	2x2 max pooling with stride [2 2] and padding "same"
18	'fc1_1'	Fully Connected	4096 fully connected layer
19	'relu5_1'	ReLU	ReLU
20	'fc1_2'	Fully Connected	4096 fully connected layer
21	'relu5_2'	ReLU	ReLU
22	'fc2'	Fully Connected	128 fully connected layer
23	'EmbeddingBatch'	ReLU	ReLU
24	'regressionoutput'	Regression Output	Mean-squared-error

TABLE III
DATASET INFORMATION

Subjects (t_0)	Subjects ($t_0 + t_1$)	Clinical score range
Level 1: 18 Healthy 37 Patient	**Level 1:** 18 Healthy 58 Patient	/
Level 2: 20 Low 17 High	**Level 2:** 33 Low 25 High	Low=[0-1] High=[2-3]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

TABLE IV
PERFORMANCE METRICS

Measure	Binary Classification	Multi-class Classification	Hierarchical Classification
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}}{l}$	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}}{l}$
Precision	$\frac{tp}{tp+fp}$	$P_\mu=\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$ $P_M=\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	$P_\downarrow=\frac{\lvert C_\downarrow^c\cap C_\downarrow^d\rvert}{\lvert C_\downarrow^c\rvert}$
Recall	$\frac{tp}{tp+fn}$	$R_\mu=\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$ $R_M=\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	$R_\downarrow=\frac{\lvert C_\downarrow^c\cap C_\downarrow^d\rvert}{\lvert C_\downarrow^d\rvert}$
F1-Score	$2\cdot\frac{Precision\cdot Recall}{Precision+Recall}$	$F1S_\mu=2\cdot\frac{P_\mu\cdot R_\mu}{P_\mu+R_\mu}$ $F1S_M=2\cdot\frac{P_M\cdot R_M}{P_M+R_M}$	$F1S_\downarrow=2\cdot\frac{P_\downarrow\cdot R_\downarrow}{P_\downarrow+R_\downarrow}$

In the first column, measures for binary classification are reported: *tp* represent the true positive, *tn* the true negative, *fp* the false positive and *fn* the false negative. In the second column, the same measures are generalized for a multi-class problem considering many classes *Ci*. μ and M are referred to micro- and macro-averaging. Finally, in the third column there are the measures for hierarchical classification: $C_\downarrow^c$ are the subclasses of C assigned by the classifier while $C_\downarrow^d$ are the labels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

TABLE V
HIERARCHICAL APPROACH PERFORMANCE METRICS [LEVEL 1: HEALTHY VS PATIENTS – LEVEL 2: LOW SEVERITY VS HIGH SEVERITY]

Measure	Machine Learning			Transfer Learning			Combination: Machine Learning (Level 1) + Transfer Learning (Level 2)		
	5-FOLD	10-FOLD	LEAVE-ONE-OUT	5-FOLD	10-FOLD	LEAVE-ONE-OUT	5-FOLD	10-FOLD	LEAVE-ONE-OUT
Accuracy	1: 88.08% 2: 61.02% Total: 57.89%	1: 86.84% 2: 51.72% Total: 51.32%	1: 88.16% 2: 60.66% Total: 56.58%	1: 78.95% 2: 75% Total: 60.53%	1: 84.11% 2: 77.42% Total: 65.79%	1: 84.21% 2: 78.33% Total: 67.11%	**1: 93.42% 2: 78.69% Total: 76.32%**	**1: 88.16% 2: 75.44% Total: 69.74%**	**1: 88.16% 2: 78.69% Total: 71.05%**
Precision	1: 91.53% 2: 58.52% Total: 73%	1: 91.38% 2: 48.95% Total: 68.48%	1: 90.16% 2: 56.37% Total: 72.37%	1: 87.50% 2: 79.18% Total: 69.96%	1: 87.10% 2: 77.81% Total: 74.73%	1: 88.33% 2: 78.70% Total: 75.66%	**1: 93.44% 2: 78.55% Total: 84.67%**	**1: 92.98% 2: 75.62% Total: 78.93%**	**1. 90.16% 2. 78.87% Total: 78.61%**
Recall	1: 93.10% 2: 55.36% Total: 73%	1: 91.38% 2: 49.02% Total: 68.48%	1: 94.83% 2: 53.66% Total: 72.37%	1: 84.48% 2: 75.52% Total: 69.96%	1: 93.10% 2: 77.81% Total: 79.28%	1: 91.33% 2: 79.86% Total: 75.66%	**1: 98.28% 2: 79.50% Total: 84.67%**	**1: 91.38% 2: 76.60% Total: 78.93%**	**1. 94.83 2. 80.24% Total: 78.61%**
F1-Score	1: 92.31% 2: 56.90% Total: 73%	1: 91.38% 2: 48.98% Total: 48.95%	1: 92.44% 2: 54.98% Total: 72.37%	1: 85.96% 2: 75.26% Total: 69.96%	1: 90% 2: 78.54% Total: 74.73%	1: 89.33% 2: 79.28% Total: 75.66%	**1: 95.80% 2: 79.02% Total: 84.67%**	**1: 92.98% 2: 75.62% Total: 78.93%**	**1. 92.44 2. 79.55% Total: 78.61%**

Detailed performance metrics of HMLM for each level (1-2) in cascade combining machine learning, transfer learning and machine learning + transfer learning respectively. Each parameter has been extracted using the ensemble majority voting technique with four classifiers (SVM, k-NN, Naïve-Bayes and Decision Tree).

TABLE VI
COMPARISON BETWEEN CLINICAL ASSESSMENT (TARGET) AND HIERARCHICAL MODEL (PREDICTED) FOR EACH SUBJECT

CLINICAL ASSESSMENT (TARGET)			HIERARCHICAL MODEL (PREDICTED)	
Presence of ATAXIA	**Speech Disturbance Severity**	**Speech Disturbance Score**	**Presence of ATAXIA**	**Speech Disturbance Severity**
Patient	High	2	Patient	High
Patient	Low	1	Patient	High
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	High
Patient	Low	1	Patient	Low
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	High
Patient	Low	0	Patient	Low
Patient	Low	1	Patient	Low
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	High
Patient	Low	1	Patient	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Healthy	Low	0	Patient	Low
Healthy	Low	0	Healthy	Low
Healthy	Low	0	Healthy	Low
Patient	Low	0	Patient	Low
Patient	Low	0	Patient	Low
Patient	Low	1	Healthy	Low
Patient	Low	1	Patient	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	High	3	Patient	High
Patient	High	3	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	0	Healthy	Low
Patient	Low	1	Patient	Low

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

Patient	High	2	Patient	High
Patient	High	2	Patient	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	High
Patient	High	2	Patient	High
Patient	High	3	Patient	Low
Patient	Low	1	Patient	High
Healthy	Low	0	Patient	Low
Healthy	Low	0	Healthy	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	High	2	Patient	Low
Patient	High	2	Patient	High
Healthy	Low	0	Healthy	Low
Patient	High	2	Patient	High
Healthy	Low	0	Healthy	Low
Patient	High	2	Patient	High
Patient	High	2	Patient	High
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Patient	High	3	Patient	High
Healthy	Low	0	Healthy	Low
Patient	High	2	Healthy	Low
Patient	Low	1	Patient	Low
Healthy	Low	0	Patient	Low
Healthy	Low	0	Patient	Low
Patient	Low	1	Patient	High
Patient	Low	1	Patient	High
Healthy	Low	0	Patient	Low
Patient	Low	1	Patient	Low
Patient	Low	1	Patient	Low
Healthy	Low	0	Patient	Low
Patient	Low	1	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	High
Healthy	Low	0	Healthy	Low
Patient	Low	1	Patient	Low

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3135078, IEEE Access

TABLE VII
FLAT MULTI-CLASS APPROACH PERFORMANCE METRICS
(SINGLE MODEL WITH THREE-CLASSES [HEALTHY, PATIENTS WITH LOW SEVERITY AND PATIENTS WITH HIGH SEVERITY])

Measure	Machine Learning			Transfer Learning		
	5-fold	10-fold	Leave-one-out	5-fold	10-fold	Leave-one-out
Accuracy	57.14%	54.23%	57.89%	65.58%	65.89%	65.79%
Precision Macro (M)	63.91%	62.43%	63.97%	66.58%	66.37%	66.05%
Precision Micro (μ)	57.14%	54.23%	57.89%	65.58%	65.89%	65.79%
Recall Macro (M)	41.24%	40.10%	41.59%	48.35%	48.71%	48.66%
Recall Micro (μ)	40%	37.20%	40.74%	48.81%	49.19%	49.02%
F1-Score Macro (M)	50.13%	48.84%	50.41%	55.83%	56.26%	56.15%
F1-Score Micro (μ)	47.06%	44.13%	47.83%	55.97%	56.33%	56.18%

Detailed performance metrics of flat multi-class approach testing machine learning and transfer learning. Each parameter has been extracted using the ensemble majority voting technique with four classifiers (SVM, k-NN, Naïve-Bayes and Decision Tree).