

Recognition of Dysarthria in Amyotrophic Lateral Sclerosis patients using Hypernetworks

Loukas Ilias, Dimitris Askounis
DSS Lab, School of ECE, NTUA
15780 Athens, Greece
{lilias,askous}@epu.ntua.gr

Abstract—Amyotrophic Lateral Sclerosis (ALS) constitutes a progressive neurodegenerative disease with varying symptoms, including decline in speech intelligibility. Existing studies, which recognize dysarthria in ALS patients by predicting the clinical standard ALSFRS-R, rely on feature extraction strategies and the design of customized convolutional neural networks followed by dense layers. However, recent studies have shown that neural networks adopting the logic of input-conditional computations enjoy a series of benefits, including faster training, better performance, and flexibility. To resolve these issues, we present the first study incorporating hypernetworks for recognizing dysarthria. Specifically, we use audio files, convert them into log-Mel spectrogram, delta, and delta-delta, and pass the resulting image through a pretrained modified AlexNet model. Finally, we use a hypernetwork, which generates weights for a target network. Experiments are conducted on a newly collected publicly available dataset, namely VOC-ALS. Results showed that the proposed approach reaches Accuracy up to 82.66% outperforming strong baselines, including multimodal fusion methods, while findings from an ablation study demonstrated the effectiveness of the introduced methodology. Overall, our approach incorporating hypernetworks obtains valuable advantages over state-of-the-art results in terms of generalization ability, parameter efficiency, and robustness.

Index Terms—amyotrophic lateral sclerosis, speech impairment, dysarthria, hypernetworks

I. INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease, which damages both upper and lower motor neurons over time. This degeneration results in dysphagia, impairment of speech intelligibility due to dysarthria, limb paralysis, and respiratory failure [1]–[4]. Studies have shown that ALS affects between 4.1 and 8.4 per 100,000 persons [5]. Although the survival time is approximately three years, approximately 20% of individuals with ALS live for five years, 10% survive for 10 years, and 5% live for 20 years or more [6]. Diagnosis of ALS is a difficult task [7], since there is no single medical test. However, early diagnosis is crucial, so as to ensure that individuals will receive treatment to slow ALS progression and maintain a good Quality of Life [8], [9]. Since ALS leads to dysarthria, researchers have proposed methods for recognizing dysarthria in ALS patients. The recognition of dysarthria is achieved by speech-language pathologists through the Revised ALS Functional Rating Scale (ALSFRS-R) [10]. Specifically, ALSFRS-R indicates severity levels of dysarthria from 0 to 4, where 0 indicates loss of useful speech, 1 denotes speech combined with nonvocal communications,

2 indicates intelligible with repeating, 3 denotes detectable speech disturbance, and 4 denotes normal speech.

Existing studies rely on feature extraction strategies followed by feature selection techniques and train of traditional Machine Learning (ML) algorithms. However, this is a tedious procedure demanding feature expertise, while it is not ensured that the optimal set of features is found. Recently, researchers extract log-Mel spectrograms or Mel-frequency Cepstral Coefficients (MFCC), their delta, and double-delta and train deep neural networks. However, these approaches use customized Convolutional Neural Networks (CNNs), obtaining often suboptimal performance, since these approaches depend on limited data. Also, these customized CNNs are followed by dense layers, which are accountable for doing everything and thus do not adopt an input-conditional computation logic. On the other hand, hypernetworks [11], [12] constitute a powerful deep learning technique, which ensures greater flexibility adaptability, faster training, information sharing, model compression, and so on. Hypernetworks have been proved advantageous for multiple tasks, including causal inference, natural language processing, transfer learning, and weight pruning. These networks are also beneficial for tasks with limited data. Specifically, hypernetworks are neural networks, which generate weights for another neural network, namely the target network. By generating weights dynamically, hypernetworks introduce implicit regularization, improving generalization performance.

To tackle the aforementioned limitations, we present the first study for recognizing dysarthria in ALS patients using hypernetworks. Specifically, we use audio files, which correspond to syllable repetitions of /pa/, and transform them into images of three channels, namely log-Mel spectrogram, delta, and delta-delta. Next, we pass each image through an AlexNet pretrained model and get an image representation vector. After that, motivated by the fact that hypernetworks ensure faster training and are beneficial for tasks with limited data, we employ a hypernetwork, which receives as input a condition vector, which follows normal distribution. This hypernetwork generates weights for the target network, which receives as input the output of AlexNet. Finally, the output layer corresponds to the prediction of dysarthria, i.e., binary classification task. Results are performed on the publicly available VOC-ALS dataset [13]. Finally, a series of ablation experiments is performed for exploring the effectiveness of the introduced

method. Results show that our proposed approach improves state-of-the-art ones, while also offering multiple advantages over existing research initiatives due to the inherent benefits of hypernetworks, i.e., parameter efficiency, task adaptation - robustness, generalization ability.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first study using hypernetworks to detect dysarthria in ALS patients.
- We compare our approach with strong baselines, including multimodal fusion methods.
- We perform a series of ablation experiments to explore the effectiveness of our approach.

II. RELATED WORK

A. Traditional Machine Learning Algorithms

Dubbioso et al. [14] extracted a set of acoustic features from different tasks, performed feature selection strategies, and trained a Decision Tree classifier for differentiating healthy subjects from non-healthy ones and predicting dysarthria severity levels in ALS patients. Experiments were performed on different tasks, including reading, monologue, and vocalization. In [15], the authors extracted a set of acoustic and articulatory features and trained Ridge regression and a Support Vector Machine to predict the ALSFRS-R score. In terms of the acoustic features, the authors used MFCC, their delta, and delta-delta and computed some statistics, e.g., mean, standard deviation. Regarding articulatory features, the authors computed a distance matrix and computed some statistics, e.g., skewness, kurtosis, and so on. The study in [16] was focused on the detection of ALS with bulbar involvement. The authors extracted a set of features, including jitter, shimmer, harmonics-to-noise ratio, pitch, and so on. Principal component analysis was used for dimensionality reduction. Finally, the authors trained the following machine learning classifiers: SVM, neural network with a hidden layer, LDA, Logistic Regression, Naive Bayes, and Random Forest (RF). Vashkevich and Rushkevich [17] proposed a study based on voice analysis to detect ALS patients. Specifically, the authors extracted a set of acoustic features from phonation vowels /a/ and /i/, performed feature selection algorithms, and trained a linear discriminant analysis for the classification purposes. The authors in [18] extracted a set of acoustic features and trained a bayesian logistic regression model for differentiating the following groups: (i) control vs ALS, (ii) control vs ALS-early, and (iii) ALS-early vs ALS-late. The main limitation of this study is related to the imbalanced dataset between ALS and control participants. Specifically, the dataset includes 119 ALS patients and 22 healthy controls.

B. Deep Neural Networks

Two different transfer learning strategies were introduced in [19]. Specifically, the authors explored fine-tuning and multitask learning frameworks. As auxiliary tasks, the authors used input feature reconstruction and gender classification. The authors used as input to the deep neural networks a vector consisting of MFCC (excluding energy coefficient) with delta

and double delta features. The deep neural network comprised a series of dense layers. Three set of experiments were performed in the study of [20], including (1) classification among ALS, Parkinson disease (PD), and Healthy control, (2) 5-class ALS severity classification based on ALSFRS-R, and (3) 3-class PD severity classification. The authors used as input log-Mel spectrograms and passed them through CNN layers followed by fully connected layers. Four tasks were used, including spontaneous speech, image description, sustained phonation, and diadochokinetic rate. In [21], the authors segmented the audio file into non-overlapping audio frames, converted it into log-Mel spectrogram and passed each frame through CNN layers. Next, they aggregated each frame's output to get the final prediction for the entire voice signal. The task was the prediction of the ALSFRS-R score. Three approaches were employed for classifying ALS patients and healthy control in [22]. In terms of the first approach, the authors extracted features using the openSMILE toolkit and trained an Artificial Neural Network with one hidden layer. Regarding the other two approaches, the authors utilized filterbank, delta, and delta-delta as input to time-CNNs and frequency-CNNs.

C. Related Work Review Findings

As is evident in Section II-A, existing studies focus on the extraction of acoustic features and the train of shallow machine learning classifiers, which constitutes a tedious procedure and does not generalize to new subjects. As is presented in Section II-B, existing studies convert the audio files into log-Mel spectrograms, delta, and delta-delta and pass them through CNN layers followed by dense layers.

Our study is different from existing studies, since we present the first study incorporating hypernetworks into a deep neural network for recognizing dysarthria in ALS patients. Also, this study has been performed in a newly collected dataset, which is publicly available.

III. DATASET AND TASK

We use the VOC-ALS dataset described in [13] to perform our experiments. VOC-ALS is a newly collected publicly available dataset and represents the most comprehensive freely downloadable dataset. Specifically, this dataset comprises 51 healthy controls and 102 ALS patients. Each participant is asked to perform a series of tasks using a smartphone application, including phonation of the vowels /a/, /e/, /i/, /o/, /u/ and /pa/, /ta/, /ka/ syllable repetition, recordings of subjects vocalizing the days of the week, reading task, and monologue. However, the authors have made publicly available only the audio files corresponding to the phonation of the vowels /a/, /e/, /i/, /o/, /u/ and /pa/, /ta/, /ka/ syllable repetition.

In this study, the task is to categorize ALS patients into dysarthric and non-dysarthric ones. According to the ALSFRS-R score, four classes are available in terms of the ALS patients, including severe, moderate, mild, and non-dysarthric. A detailed description of the number of patients and demographic information per category is reported in Table I. For performing

our experiments, we merge severe, moderate, and mild classes into one class, i.e., dysarthric.

TABLE I
DESCRIPTION OF VOC-ALS DATASET.

Severity class	Severe	Moderate	Mild	Normal
ALSFRRS-R	1	2	3	4
# M: # F	2:3	9:4	21:10	33:20
Age (M:F)	64.00:65.67	65.44:72.00	65.48:60.90	60.15:61.55
Total	5	13	31	53

IV. METHODOLOGY

In this section, we describe our introduced methodology for recognizing dysarthria in ALS patients. Below, we describe in detail each component of our proposed architecture. Our proposed approach is illustrated in Fig. 1.

Input. We use repetitions of the syllable /pa/, denoted as *rhythmPA* in the dataset. Next, we convert each audio file into log-Mel spectrogram, delta, and double-delta. To do this, we utilize the Python library, namely *librosa* [23]. In our experiments, we use 256 Mel bands, hop length accounting for 512. Let the input representation be $x \in \mathbb{R}^{3 \times 224 \times 224}$.

Deep Learning Model - AlexNet. Next, we pass x through a pretrained AlexNet model [24]. AlexNet consists of five convolutional layers followed by fully connected layers. The first convolutional layer uses large filters (11x11) to capture low-level patterns, followed by progressively smaller filters (5x5 and 3x3) to extract finer details. ReLU activation and max-pooling layers are used in AlexNet architecture.

In our experiments, we modify AlexNet by removing the last dense (fully connected) layer. Let the output of the AlexNet model be $X \in \mathbb{R}^D$, where $D = 768$.

Hypernetwork. Next, we use hypernetworks [11], [12]. Specifically, hypernetworks are neural networks which are trained for generating weights for another neural network, known as the target network. Unlike traditional models where parameters are directly optimized, a hypernetwork leverages a context vector C , which generates weights for the target network dynamically. The context vector C can be task-, data-, or noise-conditioned, allowing the target network to adapt more efficiently.

As illustrated in Fig. 1, let the hypernetwork be denoted as $H(C; \Phi)$, which maps the context vector C to the parameters Θ of the target network $F(X; \Theta)$. The output of the hypernetwork, Θ , represents the generated weights and biases, which is given to the target network. In our experiments, we set C to follow a normal distribution. Specifically, $C \in \mathbb{R}^d$, where $d = 128$. Thus, the output of the AlexNet, denoted as $X \in \mathbb{R}^D$, where $D = 768$, is given as input to the target network, $F(X; \Theta)$, where Θ denotes the parameters learnt by $H(C; \Phi)$. Formally:

$$\Theta = H(C; \Phi) \quad (1)$$

$$\hat{y} = F(X; \Theta) \quad (2)$$

The hypernetwork consists of fully connected layers and ReLU activation functions transforming into a meaningful

weight representation. In terms of the implementation details, the hypernetwork includes a hidden layer with 512 units and an output layer of 1536 units. The hypernetwork outputs a weight matrix $W \in \mathbb{R}^{768 \times 2}$ and a bias term $b \in \mathbb{R}^2$. The target network performs a matrix multiplication between the extracted features (X) and the generated weight matrix. The bias term is added.

Output. Let $\hat{y} \in \mathbb{R}^2$ denote the output of the target network ($F(X; \Theta)$), since our task corresponds to a binary classification task. We minimize the cross-entropy loss function.

V. EXPERIMENTS AND RESULTS

A. Baselines

- /pa/ + eGeMAPS + Random Forest: This method utilizes the audio files corresponding to the syllable repetition /pa/. Then, this method uses the openSMILE toolkit [25] and extracts the eGeMAPS feature set (88d). Then, a Random Forest classifier is trained.
- vocalization of the days of the week, reading task, and monologue features + Random Forest: This method utilizes the features provided by the authors in [13], including mean and std F0, jitter, shimmer, etc., and trains an RF classifier.
- Introduced approach with inputs the phonations of vowels and syllable repetitions: This method uses as inputs to the proposed deep neural network described in Section IV speech signals corresponding to the phonation of vowels /a/, /e/, /i/, /o/, /u/, and the syllable repetitions of /ta/ and /ka/.
- Multimodal Fusion Method with input /pa/ and /ta/: This method uses as input speech signals corresponding to the /pa/ and /ta/ syllable repetitions. The audio signals are converted into log-Mel spectrogram, delta, and delta-delta and are given as input to pretrained AlexNet models sharing the same weights. The output vectors of the AlexNet models (768d) are fed into a fusion method, namely Gated Multimodal Unit (GMU) [26]. Let f^t and f^v denote the /pa/ and /ta/ representation vectors respectively. The equations governing the GMU are described as follows: $h^t = \tanh(W^t f^t + b^t)$, $h^v = \tanh(W^v f^v + b^v)$, $z = \sigma(W^z [f^v; f^t] + b^z)$, $h = z * h^v + (1 - z) * h^t$, $\Theta = \{W^t, W^v, W^z\}$, where Θ denote the learnable parameters, and $[:, :]$ the concatenation operation. σ is the sigmoid activation function. h is the output of the GMU. The output of GMU is passed through a dense layer with two units.
- Concatenation of all speech signals: This method uses as input the speech signals corresponding to the phonation of the vowels /a/, /e/, /i/, /o/, /u/ and /pa/, /ta/, /ka/ syllable repetition. Firstly, speech signals are converted into log-Mel spectrogram, delta, and delta-delta. Secondly, they are given as input to pretrained AlexNet models sharing the same weights. The output vectors of the AlexNet models (32d) are concatenated into one vector. The resulting vector (256d) is passed through a dense layer of two units.

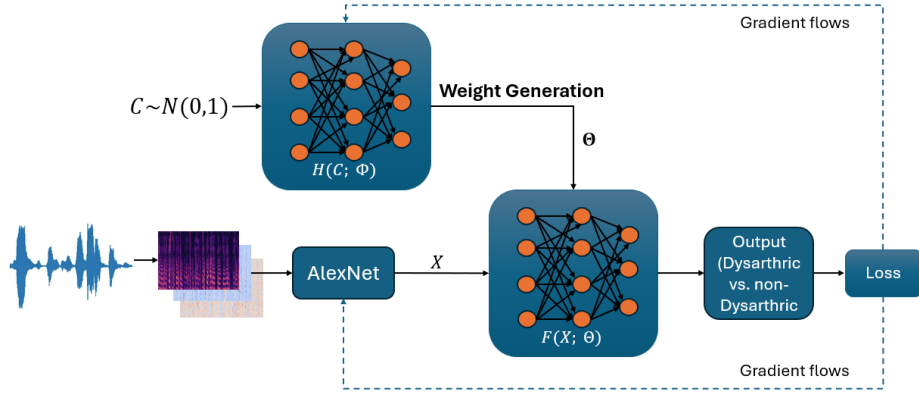


Fig. 1. Illustration of our proposed methodology. Each speech signal is transformed into log-Mel spectrogram, delta, and delta-delta, and is given as input to a pretrained AlexNet model. The output vector of the AlexNet model with a dimensionality of 768 is given as input to a target network ($F(X; \Theta)$), where its weights are generated by a hypernetwork ($H(C; \Phi)$). The input to the hypernetwork is denoted by C and follows a normal distribution. Finally, we use an output layer consisting of two units, which differentiates dysarthric from non-dysarthric ALS patients.

B. Experimental Setup

We use a 5-fold cross-validation framework with four repetitions to train and test our proposed model. We train the proposed deep learning model for 30 epochs. We use PyTorch for performing our experiments. We use a learning rate of $1e-5$. Experiments are conducted on a single NVIDIA A100 PCIe 80GB GPU.

C. Evaluation Metrics

Accuracy, Precision, Recall, F1-score, and Specificity are used to evaluate our introduced approach. We report the mean and standard deviation of these metrics obtained over 10 runs in a 5-fold cross-validation scheme.

D. Results

Table II reports the results of our proposed approach. Specifically, this table presents a comparison of our approach with the baselines described in Section V-A. As one can observe, our introduced model surpasses the baselines in Recall by 0.83-31.78%, in F1-score by 2.72-28.95%, and in Accuracy by 3.02-22.61%. We observe that the proposed approach with the input of the vowel /u/ yields the highest Specificity accounting for 89.59%. However, F1-score is a more important metric than Specificity, since a lower F1-score means that dysarthric ALS patients are misdiagnosed as non-dysarthric ones. As one can observe in Table II, syllable repetitions of /pa/, /ka/, and /ta/ perform better than the vowels. We also observe that the vowel /u/ presents the best performance among the other vowels. Additionally, the train of a Random Forest classifier utilizing the eGeMAPS feature set of /pa/ yields the lowest evaluation metrics. The fusion of the input representations of /pa/ and /ta/ via a Gated Multimodal Unit presents a decrease in Accuracy in comparison with the unimodal models. Specifically, *Fusion (/pa+/ta/)* presents a decrease in Accuracy in comparison with /pa/ (5.42%) and /ta/ (3.02%). Finally, the fusion method of concatenation yields lower Accuracy and F1-score compared with most vowels or syllables. We hypothesize that these

differences in performances are attributable to the fact that concatenation does not capture the inherent correlations of the input representations.

TABLE II
PERFORMANCE COMPARISON AMONG PROPOSED MODELS AND BASELINES. BEST RESULTS PER EVALUATION METRIC ARE IN BOLD.

Architecture	Evaluation metrics				
	Precision	Recall	F1-score	Accuracy	Specificity
Comparison with baselines					
/pa/+eGeMAPS+RF	63.62 ±13.99	55.00 ±16.12	58.27 ±13.78	63.23 ±10.72	70.86 ±12.84
weekdays,reading, monologue features+RF	62.45 ±13.50	48.44 ±17.96	52.46 ±12.41	60.05 ±8.52	71.09 ±14.77
/ta/	80.13 ±9.12	79.39 ±11.34	78.69 ±5.35	79.64 ±4.50	79.45 ±11.63
/ka/	81.78 ±12.42	76.94 ±6.63	78.60 ±6.54	79.61 ±7.28	82.09 ±14.14
/a/	82.94 ±11.96	67.77 ±16.16	72.75 ±10.57	76.74 ±6.97	85.14 ±11.79
/e/	78.76 ±11.16	66.72 ±13.71	70.98 ±8.42	74.43 ±6.21	81.45 ±12.78
/i/	77.62 ±10.35	74.94 ±16.16	75.26 ±10.77	77.19 ±8.10	79.23 ±10.25
/o/	80.00 ±9.71	72.99 ±10.63	75.35 ±5.27	77.24 ±4.79	81.18 ±11.76
/u/	86.65 ±9.44	66.44 ±12.12	74.35 ±8.45	78.45 ±6.43	89.59 ±8.47
Fusion (/pa+/ta/)	76.96 ±10.70	75.66 ±14.82	75.70 ±11.06	77.21 ±9.69	78.77 ±10.25
Concatenation	77.75 ±12.39	72.94 ±13.26	74.40 ±9.93	76.05 ±9.31	78.82 ±14.90
Introduced Approach					
/pa/	84.18 ±7.99	80.22 ±11.15	81.41 ±6.10	82.66 ±5.44	84.73 ±9.34

E. Ablation Study

In this section, we perform a series of ablation experiments to investigate the effectiveness of the proposed approach. Firstly, we replace the hypernetwork with a simple dense layer. Therefore, the output vector of the AlexNet model is fed to a dense layer of two units. Findings showed that Accuracy and F1-score presented a decrease of 2.46% and 2.08% respectively. Secondly, we use a data-conditioned hypernetwork. Specifically, we use the eGeMAPS feature set (88d) as the condition vector, i.e., input to the hypernetwork $H(C; \Phi)$, instead of a vector following a normal distribution. Results showed that Accuracy and F1-score presented a decline of

4.22% and 4.98% respectively. Thirdly, we use as input MFCC instead of log-Mel spectrogram features. Therefore, the input to the deep neural network is the transformation of the speech signal into MFCC, delta, and delta-delta. Results indicated that Accuracy and F1-score dropped by 2.48% and 3.69% respectively. Finally, we do not use pretrained AlexNet model. Specifically, we use AlexNet with no pretrained weights. Findings showed that Accuracy and F1-score had a decrease of 3.33% and 4.53% respectively.

TABLE III
ABLATION STUDY. BEST RESULTS PER EVALUATION METRIC ARE IN BOLD.

Architecture	Evaluation metrics				
	Precision	Recall	F1-score	Accuracy	Specificity
Ablation Experiments					
<i>Removal of Hypernetwork</i>	81.25 ±11.66	78.66 ±9.31	79.33 ±7.62	80.20 ±7.88	81.68 ±12.18
<i>Usage of eGeMAPS as condition vector</i>	80.07 ±10.86	74.88 ±13.45	76.43 ±9.29	78.44 ±7.68	81.68 ±10.73
<i>MFCC</i>	85.48 ±10.75	73.50 ±13.50	77.72 ±8.26	80.18 ±6.79	86.14 ±12.10
<i>AlexNet with no pretrained weights</i>	82.48 ±10.00	73.94 ±14.88	76.88 ±9.52	79.33 ±7.69	84.50 ±9.02
Introduced Approach					
	84.18 ±7.99	80.22 ±11.15	81.41 ±6.10	82.66 ±5.44	84.73 ±9.34

VI. CONCLUSION

In this paper, we present the first study integrating hypernetworks into a deep neural network for identifying dysarthria in ALS patients. Specifically, after converting each audio signal into an image of three channels, namely log-Mel spectrogram, delta, and delta-delta, we pass the resulting image into AlexNet. Then, a hypernetwork is used for producing weights for the target network. Specifically, the output vector of AlexNet is given as input to the target network, while a vector with normal distribution is given as input to the hypernetwork. Results showed that the proposed approach yielded an Accuracy of 82.66%, while results of an ablation study demonstrated the effectiveness of the introduced approach. The VOC-ALS dataset is imbalanced in terms of the severity levels of dysarthria in ALS patients. For this reason, in this study, we did not experiment with predicting the severity level of dysarthria. In the future, we aim to use neural architecture search approaches for finding the optimal architecture in our task.

REFERENCES

- [1] K. Joubert, J. Bornman, and E. Alant, "Speech intelligibility and marital communication in amyotrophic lateral sclerosis: an exploratory study," *Communication Disorders Quarterly*, vol. 33, no. 1, pp. 34–41, 2011.
- [2] A. Chiò, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, and O. B. of the Eurals Consortium, "Prognostic factors in als: A critical review," *Amyotrophic Lateral Sclerosis*, vol. 10, no. 5-6, pp. 310–323, 2009.
- [3] A. Al-Chalabi and O. Hardiman, "The epidemiology of als: a conspiracy of genes, environment and time," *Nature Reviews Neurology*, 2013.
- [4] L. Xu, T. Liu, L. Liu, X. Yao, L. Chen, D. Fan, S. Zhan, and S. Wang, "Global variation in prevalence and incidence of amyotrophic lateral sclerosis: a systematic review and meta-analysis," *Journal of neurology*, vol. 267, pp. 944–953, 2020.
- [5] E. Longinetti and F. Fang, "Epidemiology of amyotrophic lateral sclerosis: an update of recent literature," *Current opinion in neurology*, vol. 32, no. 5, pp. 771–776, 2019.

- [6] A. Association, "Stages of ALS," Available online at: <https://www.als.org/understanding-als/stages>, accessed: 2025-01-15.
- [7] S. Paganoni, E. A. Macklin, A. Lee, A. Murphy, J. Chang, A. Zipf, M. Cudkowicz, and N. Atassi, "Diagnostic timelines and delays in diagnosing amyotrophic lateral sclerosis (als)," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 15, no. 5-6, pp. 453–456, 2014.
- [8] A. Chiò, A. Gauthier, A. Montuschi, A. Calvo, N. Di Vito, P. Ghiglione, and R. Mutani, "A cross sectional study on determinants of quality of life in als," *Journal of Neurology, Neurosurgery & Psychiatry*.
- [9] H. for Special Surgery (HSS), "ALS (Amyotrophic Lateral Sclerosis)," Available online at: https://www.hss.edu/condition-list_amyotrophic-lateral-sclerosis.asp, accessed: 2025-01-15.
- [10] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, no. 1, pp. 13–21, 1999.
- [11] D. Ha, A. M. Dai, and Q. V. Le, "Hypernetworks," in *International Conference on Learning Representations*, 2017. [Online]. Available: <https://openreview.net/forum?id=rkpACe1lx>
- [12] V. K. Chauhan, J. Zhou, P. Lu, S. Molaei, and D. A. Clifton, "A brief review of hypernetworks in deep learning," *Artificial Intelligence Review*, vol. 57, no. 9, p. 250, 2024.
- [13] R. Dubbioso, M. Spisto, L. Verde, V. V. Iuzzolino, G. Senerchia, E. Salvatore, G. De Pietro, I. De Falco, and G. Sannino, "Voice signals database of als patients with different dysarthria severity and healthy controls," *Scientific Data*, vol. 11, no. 1, p. 800, 2024.
- [14] R. Dubbioso, M. Spisto, L. Verde, V. V. Iuzzolino, G. Senerchia, G. De Pietro, I. De Falco, and G. Sannino, "Precision medicine in als: Identification of new acoustic markers for dysarthria severity assessment," *Biomedical Signal Processing and Control*, vol. 89, p. 105706, 2024.
- [15] A. Wisler, K. Teplansky, J. Green, Y. Yunusova, T. Campbell, D. Heitzman, and J. Wang, "Speech-based estimation of bulbar regression in amyotrophic lateral sclerosis," in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*. Minneapolis, Minnesota: ACL, Jun. 2019.
- [16] A. Tena, F. Claria, F. Solsona, E. Meister, and M. Povedano, "Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: Diagnostic decision support development study," *JMIR Med Inform*, 2021.
- [17] M. Vashkevich and Y. Rushkevich, "Classification of als patients based on acoustic analysis of sustained vowel phonations," *Biomedical Signal Processing and Control*, vol. 65, p. 102350, 2021.
- [18] L. E. Simmat, J. Robin, M. J. Spilka, and Y. Yunusova, "Detecting bulbar amyotrophic lateral sclerosis (als) using automatic acoustic analysis," *BioMedical Engineering OnLine*, vol. 23, no. 1, p. 15, 2024.
- [19] T. Bhattacharjee, A. Jayakumar, Y. Belur, A. Nalini, R. Yadav, and P. K. Ghosh, "Transfer learning to aid dysarthria severity classification for patients with amyotrophic lateral sclerosis," in *INTERSPEECH 2023*.
- [20] B. Suhas, J. Mallela, A. Illa, B. Yamini, N. Atchayaram, R. Yadav, D. Gope, and P. K. Ghosh, "Speech task based automatic classification of als and parkinson's disease and their severity using log mel spectrograms," in *2020 International Conference on Signal Processing and Communications (SPCOM)*, 2020, pp. 1–5.
- [21] F. G. Vieira, S. Venugopalan, A. S. Premasiri, M. McNally, A. Jansen, K. McCloskey, M. P. Brenner, and S. Perrin, "A machine-learning based objective measure for als disease severity," *NPJ digital medicine*, 2022.
- [22] K. An, M. Kim, K. Teplansky, J. Green, T. Campbell, Y. Yunusova, D. Heitzman, and J. Wang, "Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks," in *Interspeech 2018*, 2018, pp. 1913–1917.
- [23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015.
- [24] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.
- [25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010.
- [26] J. Arevalo, T. Solorio, M. Montes-y Gomez, and F. A. González, "Gated multimodal networks," *Neural Computing and Applications*, pp. 1–20, 2020.