

RESEARCH

Open Access



Domain generalization for voice-based cognitive impairment detection

Minsoo Kim¹, Young Chul Youn^{2,3*}, Yugwon Won¹, Hyunjoo Choi⁴, YongSoo Shim⁵, Nayoung Ryoo⁵, Ho Tae Jeong⁶, Gihyun Yun¹, Hunboc Lee¹ and SangYun Kim⁷

Abstract

Background Voice biomarkers hold potential for early cognitive disorder detection, but variations in recording conditions across different environments present challenges for accurate diagnosis using artificial intelligence (AI) models. This study aims to develop a robust, generalizable model for reliably diagnosing cognitive impairments across varied datasets.

Methods We implemented a domain generalization approach using an adapted Deep Domain-Adversarial Image Generation (DDAIG) framework. This method transforms input data to reduce center-specific characteristics and emphasizes domain-invariant features, allowing the model to focus on cognitive impairment indicators.

Results Before applying domain generalization, both cognitive impairment (CI) and center classification models achieved accuracies of 0.96. After implementing domain generalization, the CI classification accuracy decreased to 0.90, while the center classification model's accuracy dropped to 0.64. This reduction in the center classification metrics reflects the model's reduced dependence on center-specific features, indicating effective domain generalization.

Conclusion The adapted DDAIG framework effectively reduced center-specific learning, enhancing the model's ability to generalize cognitive impairment classifications across different centers. These findings suggest the role of domain generalization in developing reliable AI diagnostic tools for cognitive disorder detection.

Keywords Voice, Biomarker, Machine learning, Domain generalization, Alzheimer's disease, Dementia

*Correspondence:

Young Chul Youn
neudoc@cau.ac.kr

¹Research and Development, Baikal AI Inc., Seoul, Republic of Korea

²Department of Neurology, Chung-Ang University College of Medicine, Seoul, Republic of Korea

³Department of Medical Informatics, Chung-Ang University College of Medicine, Seoul, Republic of Korea

⁴Department of Communication Disorders, Korea Nazarene University, Cheonan, Republic of Korea

⁵Department of Neurology, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

⁶Department of Neurology, Chung-Ang University Hospital, Seoul, Republic of Korea

⁷Department of Neurology, Seoul National University College of Medicine, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, Republic of Korea

Background

The advent of digital biomarkers has generated considerable interest in the rapidly increasing field of biomarker research for cognitive disorders. Among these, vocal attributes offer promising characteristics for the early detection and prediction of mild cognitive impairments (MCIs) and dementia [1–4]. Studies such as Mahon & Lachman (2022) and Park et al. (2024) highlight how vocal changes correlate with cognitive decline, particularly in early stages of dementia [1, 4]. Thomas et al. (2020) demonstrated the utility of language and voice biomarkers in large-scale studies like the Framingham Heart Study, suggesting the feasibility of using voice data



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

to assess cognitive health remotely [2]. The advantage of vocal biomarkers lies in their non-invasive nature and potential for seamless integration into routine clinical assessments or even remote monitoring scenarios. We aimed to leverage the features of speech using machine-learning methodologies to predict cognitive impairments in various environments and domains.

Voice constitutes a complex blend of cognitive activities that are intricately connected with language, emotions, and muscle control [5–8]. The potential of voice as a biomarker arises from its sensitivity to subtle shifts in these cognitive domains, which are frequently affected in the early stages of neurodegenerative diseases [1, 9, 10]. For example, changes in tempo, pitch, clarity, and pattern of vocal errors can act as early indicators of cognitive decline. Machine learning methods can systematically analyze these vocal characteristics to detect patterns that may be too subtle for the human ear to recognize [4, 11–13]. Garcia-Gutierrez et al. (2023) and Taptiklis et al. (2023) explored machine learning's role in identifying amyloid status and predicting cognitive load from voice, respectively [11, 12].

However, using voice for the diagnosis and prognosis of cognitive disorders has its challenges. A primary issue are the variations in speech-recording conditions, as the environment and ambient noise can significantly influence the quality of speech data. In the field of artificial intelligence (AI) research focused on diagnosing and classifying various cognitive disorders [14, 15], the acquired data can be contaminated by its source, potentially skewing the results, as noted by Parsapoor et al. (2023), where variations in recording conditions impacted model accuracy [13]. This phenomenon, called the *domain shift* problem, is well documented in the literature [16–20], wherein algorithms trained on data on a single center have often exhibited lower accuracy when employed using data from different centers. Although a study attempted to predict Parkinson's disease using voice analysis [18], the model was trained on limited datasets and struggled to generalize and performed poorly on other datasets. The authors suggested that additional research is required to improve their performance such that they can be employed in real-world applications. Another study done by Zech et al. (2018) highlighted a similar concern, wherein machine learning models trained on medical imaging data from one hospital system were less effective on data from another system [19]. Such discrepancies can potentially distort the ability of the machine learning model to generalize to real-world scenarios after training on limited data. This issue of site-specific data variance presents significant challenges to the generalizability of AI diagnostic tools [20]. These findings underscore the importance of developing algorithms that are robust against diverse data collection environments to

ensure the reliability and accuracy of AI-driven diagnostics for cognitive disorders.

To discern cognitive impairment (CI) from normal cognition (NC) through machine learning, this study addresses the pivotal challenge that arises from the heterogeneous distribution of participants across various collection centers. This heterogeneity presents a significant confounding factor, as the machine learning model may inadvertently learn to identify the data source rather than the intended cognitive features associated with CI or NC. This can have profound implications, potentially resulting in a model that is more adept at center identification than accurate clinical diagnoses.

To address this issue, we explored the concept of “domain generalization,” which attempts to create models that can perform consistently well across different unseen domains. This is particularly crucial in medical settings, where data are often siloed, and the ability to generalize across various data is essential for creating robust and scalable diagnostic tools.

Domain generalization aims to mitigate the effects of domain-specific factors that can skew the performance of predictive models. In this study, the domains are represented by the centers from which the data are collected. Each center may exhibit unique characteristics in terms of ambient environmental conditions during data collection. These factors can imprint on the data, manifesting as noise that can mislead the algorithm learning.

Our approach to overcome this obstacle begins with a critical analysis of data distribution. We acknowledge that, ideally, data from NC and CI participants should be evenly collected across all centers; however, they deviate substantially. Specifically, data of patients with NC were exclusively collected from Center A, whereas those of patients with CI were collected from Centers C, E, and S. This discrepancy inadvertently transformed the task of classifying patients with NC and CI into distinguishing between Centers A and C, E, and S. Such an overlap between clinical and non-clinical factors is common in medical datasets, making it a pressing issue that must be addressed for predictive modeling.

Unlike other studies [13, 16], our proposed methodology involves a sophisticated application of denoising autoencoders, which are adept at reconstructing clean inputs from corrupted or noisy data [21]. These autoencoders are not only pivotal for feature learning but also play a crucial role in data denoising [22]. We aimed to extend this concept by not only denoising the data but also separating domain-specific characteristics, akin to separating “content” from “style” in artistic image generation [23–25]. In this context, “content” is interpreted as the cognitive-impairment features inherent to the participant, whereas “style” pertains to the center-specific characteristics. Through this approach, we expected to

develop a model that is not only resilient to domain-specific noise but also adept at extracting and leveraging the underlying cognitive features for accurate classification.

The objective of this study is to develop a robust and generalizable model that can accurately classify CIs across diverse clinical environments by implementing domain generalization techniques. This involves using denoising autoencoders to separate cognitive features from center-specific characteristics, thereby mitigating the ‘domain shift’ problem and enhancing the model’s applicability in real-world medical settings.

Methods

This study was designed to evaluate the impact of domain generalization on the ability of a neural network to classify CIs and identify center-specific characteristics. The model performance was assessed both before and after the application of domain-generalization techniques.

Data acquisition

The data used in this study were collected from Bundang Seoul National University Hospital, Anseong City, Eunpyeong St. Mary’s Hospital, and Chung-Ang University Hospital. Participants were recruited among patients with memory problems who visited the neurology clinics in these hospitals. This study was conducted in accordance with the International Conference on Harmonization Good Clinical Practice Guidelines [26]. The institutional review board (IRB) of each hospital approved the study protocol and consent forms prior to the study commencement (Seoul National University Bundang Hospital IRB No: B-2109-709-303; Anseong City, IRB-Korean Nazarene University-2021-16; Eunpyeong St. Mary’s Hospital: PC21ESDE0126; Chung-Ang University Hospital: 2110-028-481). Written informed consent was obtained from all participants before enrollment. During the task, voice recordings were obtained from the participants for analysis. This multisite data collection approach was employed to obtain a diverse range of vocal biomarkers required for developing an accurate AI algorithm capable of diagnosing Alzheimer’s disease (AD).

Inclusion criteria

The study population comprised both male and female participants aged between 60 and 79 years and were literate. Individuals with acute or severe asthma, unstable cardiovascular diseases, active peptic ulcer disease, severe liver disease, kidney issues requiring dialysis, neurological brain diseases other than AD (e.g., stroke, psychiatric disorders, epilepsy), as well as those with a history of drug or alcohol addiction in the past ten years were excluded.

Participants with CI referred to individuals with MCI or early-stage AD. The control group comprised cognitively healthy community-dwelling individuals without complaints regarding memory or motor abilities. These participants met the Christensen’s health screening criteria [27] for normal aging and scored more than one standard deviation above the mean for their age and education level in the screening tests. Additionally, they required scores of ≤ 0.42 and ≤ 6 on the Korean version of the Instrumental Activities of Daily Living Scale (K-IADL) [28] and Korean Dementia Screening Questionnaire (KDSQ) [29], respectively.

The participants were distributed across four centers: A, C, E, and S. Each participant was categorized into one of three groups: NC, MCI, and Alzheimer’s disease dementia (ADD). The patient distribution by sex (male/female) in each category across the different centers was as follows: Center A included 206 patients with NC (62 males and 144 females), Center C included 82 patients with MCI (26 males and 56 females) and 55 patients with ADD (18 males and 37 females), Center E included 82 patients with MCI (21 males and 61 females) and 24 patients with ADD (4 males and 20 females), and Center S included 87 patients with MCI (27 males and 60 females) and 31 with ADD (14 males and 17 females). The age distribution of this study subjects is as follows: Among the 110 participants diagnosed with ADD, 25 were in their 60s, and 85 were in their 70s. For the 251 participants with MCI, 63 were in their 60s, and 188 were in their 70s. The 206 NC participants included 86 individuals in their 60s and 120 individuals in their 70s.

Table 1 reflects the distribution of pre- and post-propensity score-matched participants across all centers.

Table 1 Number of pre- and post-propensity score-matched participants distributed across four centers

Center	Pre-propensity score-matched participants			Matched female participants	
	NC (M/F)	MCI (M/F)	ADD (M/F)	NC (F)	CI (F)
A	206 (62/144)			144	
C		82 (26/56)	55 (18/37)		17
E		82 (21/61)	24 (4/20)		56
S		87 (27/60)	31 (14/17)		71
Total	206 (62/144)	251 (74/177)	110 (36/74)	144	144

NC, normal cognition; MCI, mild cognitive impairment; ADD, Alzheimer’s disease dementia; CI, cognitive impairment; M, male; F, female

Table 2 Voice recording tasks

No.	Category	Task
1	Semantic verbal fluency	Naming animals
2	Reading	Reading a passage titled “Rainy Day”
3	Reading	Reading a passage titled “Birth of Eldest Son”
4	Picture description	Describing an image of a city park
5	Storytelling	Telling the fairytale “Heungbu and Nolbu”
6	Spontaneous speech	Discussing a trip to Jeju Island
7	Spontaneous speech	Explaining how to cook ramen

We used data from only female participants for training to mitigate the potential impact of gender differences on the classification results. Additionally, focusing on female data helped minimize the complexity of analysis arising by including data of both sexes. Subsequently, we performed propensity score matching to balance the data classes of female participants (Table 1). The data were randomly divided into training (80%) and validation (20%) sets, and we ensured that the same participant was not included in both sets.

Data collection

A dedicated application was developed for voice recording and used in this study. This data collection application was operated on the same model of and Android tablet (Samsung Galaxy Tab A7 SM-T500, South Korea). Participants’ voices were recorded and stored in the wav format at a sampling rate of 16 kHz in mono channel.

The tasks performed during the voice recordings were divided into seven categories, as presented in Table 2. Each task was designed to elicit various aspects of speech and language abilities that could be potentially affected by AD or other forms of CI. These tasks ranged from simple vocal activities, such as naming animals (semantic verbal fluency) and reading standardized passages (Tasks 2 and 3), and complex cognitive tasks, such as picture

descriptions (Task 4), storytelling (Task 5), and engaging in spontaneous speech regarding specific topics (Tasks 6 and 7).

The recordings captured both the interviewer’s and interviewee’s voices simultaneously. In subsequent processing, the interviewer’s voice was removed and only that of the interviewee was preserved, resulting in audio segments containing only the voice responses of interviewees. These processed audio segments served as primary data for further analysis.

The human voice generally falls within the frequency range of 300–3400 Hz. Therefore, we opted for a sampling rate of 16,000 Hz—over twice that required by Nyquist’s theorem [30]—to capture these higher frequency components and discern subtle changes in speech patterns.

Model

We employed two neural network models with distinct purposes: one for classifying CI, and the other for center classification. To ensure the model’s robustness across various domains and to avoid the overfitting of center-specific characteristics, we implemented a domain-generalization technique. In the domain generalization process, we intentionally degrade the performance of the center classifier. By doing so, the model focuses on generalizable features related to cognitive impairment rather than characteristics tied to specific recording centers.

Feature extraction (Figure. 1)

The feature extractor had an autoencoder structure; however, unlike typical autoencoders, it contained two encoders and therefore, resulted in two encodings. We named these content and style encoders and their outputs as encoded content and style, respectively. They are mathematically expressed as.

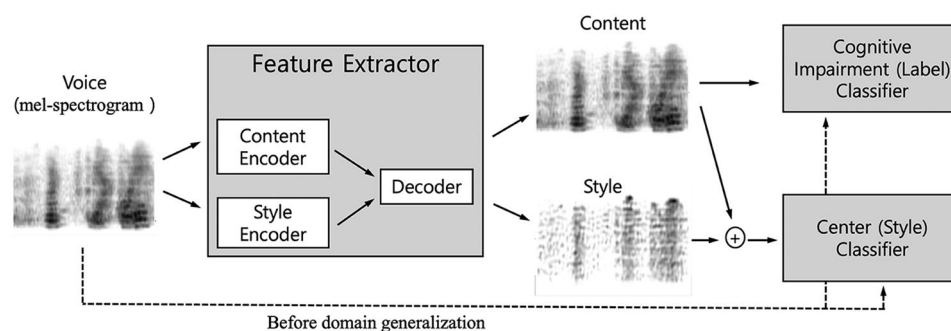


Fig. 1 Feature extraction and classification model for domain generalization. The model processes input voice data in the form of a mel-spectrogram. The Feature Extractor consists of two separate encoders: the Content Encoder, which isolates cognitive impairment features, and the Style Encoder, which captures center-specific characteristics. The outputs from these encoders are decoded and then used for classification. The Cognitive Impairment (Label) Classifier uses the content features to classify cognitive impairment, while the Center (Style) Classifier uses the combined style features to classify the center

$$\min_{\theta_s, \theta_c} L_{MSE}(g_c(x) + m_s(x), x) \quad (1)$$

where x represents the input, g_c is the content extractor with parameter θ_c (comprising the content encoder and decoder), and m_s is the style extractor with parameter θ_s (comprising the style encoder and decoder).

Both encodings are fed into the same decoder to obtain the content and style, with their sum defined as the reconstructed output. This relationship can be expressed as

$$\tilde{x} \rightarrow g_c(x) + m_s(x) \quad (2)$$

where x and $\tilde{x}$ represent the input and reconstructed output, respectively.

Both the content and style encoders comprised three residual blocks, each with two convolution layers of 3×3 kernels. Max pooling with a pool size of 2 was performed between each residual block. The convolution layers of each residual block comprised (256, 256), (128, 128), and (64, 64) channels. Additionally, the decoder comprised three residual blocks, each with two convolution layers of 3×3 kernels, but with an upsampling of size 2 between the blocks and the (64, 64), (128, 128), and (256, 256) channels.

Label classifier (CI classifier)

Label loss calculates the cross-entropy between the labels predicted by the label classifier and actual class labels. The label classifier uses content from the feature extractor to predict class labels and is mathematically expressed as.

$$\min_{\theta_l} L_l(f_l(g_c(x)), y) \quad (3)$$

where f_l denotes the label classifier parameterized by θ_l , y represents the actual class label, and L_l is the cross-entropy loss of the label classifier. The label classifier employed the Audio Spectrogram Transformer (AST) model [31, 32]. The pretrained model used in this study is the first model listed in the author's GitHub repository, titled "Full AudioSet, 10 tstride, 10 fstride, with Weight Averaging (0.459 mAP)." The label classifier, which is a binary classifier, received content as an input to determine NC or CI.

Domain classifier (center classifier)

The domain loss function calculates the cross-entropy between the domain predictions made by the domain classifier and actual domain labels. The domain classifier used the reconstructed output as the input and employed the pretrained AST model as the label classifier. It

performed a multiclass classification task to distinguish between the four centers.

$$\min_{\theta_d} L_D(h_D(g_c(x) + m_s(x)), d) \quad (4)$$

where h_D is the domain classifier with parameters θ_d , d is the domain label, and L_D represents the cross-entropy loss for the domain classifier.

Domain generalization

To mitigate unintentional learning of center characteristics by the CI classification model, we applied a domain-generalization transformation to the input data. This technique aims to render center-specific features as non-identifiable, forcing the model to focus on patterns that are more generalizable.

We employed an adapted version of the Deep Domain-Adversarial Image Generation (DDAIG) framework [33], which offers significant benefits in minimizing the influence of center-specific characteristics, which can skew model performance. The DDAIG framework consists of three fundamental components: the feature extractor, the cognition classifier, and the center classifier.

The feature extractor in our model (Fig. 1) is designed to decompose input data into two separate representations: content and style. It consists of two main components—the Content Encoder, which isolates cognitive impairment features, and the Style Encoder, which captures environment-specific characteristics. The input is a mel-spectrogram that represents audio as time-frequency data, capturing vocal features related to cognitive impairment. The Content Encoder extracts features associated with cognitive decline (e.g., vocal tempo, pitch, clarity) while excluding environmental characteristics, ensuring that the model focuses on domain-invariant cognitive features. The Style Encoder, on the other hand, captures environmental details like background noise and reverberation, isolating these from the cognitive features. This separation ensures that environmental variations do not impact cognitive classification. Both encoded outputs are combined in a Decoder to reconstruct the original input, validating that each encoder accurately isolates its designated information type. The resulting content and style outputs maintain distinct cognitive and environmental features, producing a reconstructed spectrogram that closely resembles the input.

This architecture allows the model to focus on cognitive impairment indicators while reducing environmental noise influence, enhancing robustness and generalizability across different recording settings. By isolating cognitive from environment-specific features, this approach contributes to a reliable diagnostic model resilient to recording variations.

Domain-invariant representation loss

This loss ensures that the content extractor generates features that do not contain domain-specific information, which makes it difficult for the domain classifier to accurately predict domains. The negative sign implies that during training, the direction of the gradient is reversed (gradient reversal layer) [34, 35], which essentially means that instead of minimizing the domain classifier's accuracy, it is maximized (hence, the negative sign). This is mathematically expressed as.

$$\min -L_D(h_D(gc(x)), d) \quad (5)$$

Negative loss implies that the optimization process aims to increase the domain loss, effectively making the task of the domain classifier more challenging and encouraging the model to learn domain-invariant features.

Processing

The process begins with the collection of input data from multiple centers. This data incorporates a wide variety of factors, such as the different environmental conditions at each center, which can introduce variability. To address this, initial processing is performed to normalize the features.

Input preprocessing and dimension transformation

The input data undergoes meticulous preprocessing and dimension transformation to meet the model's requirements. In this study, the input data is given in a one-dimensional form ([BATCH, SEQUENCE]), and it is transformed into a format that the model can process.

First, the dimension of the input data is set. In the case of one-dimensional input data, it is transformed into a higher-dimensional feature space and expanded to include a channel dimension, resulting in the shape [BATCH, 1, TIME, FREQUENCY].

In the feature extraction stage, a ResNet-based encoder is used to process the input data. During this process, the input data is separated into two main features: content and style. Content reflects the features necessary for CI classification, while style reflects center-specific characteristics. The extracted content and style features are then combined to reconstruct the original data. This reconstruction process maintains the main features necessary for CI classification while minimizing domain-specific biases.

For processing the output data, the output data is initially given in the form [BATCH, CHANNEL, TIME, FREQUENCY]. It is then averaged to transform it into the shape [BATCH, TIME, FREQUENCY]. Subsequently, the output data is padded to the shape [BATCH, 1024, 128] to match the dimensions expected by the model.

Specifically, the time dimension of the output data becomes 1024, and the feature dimension becomes 128.

Through this preprocessing and dimension transformation process, the model minimizes the domain-specific characteristics of the input data and learns more generalized patterns. This is crucial for the model to perform reliably across various domains and plays a significant role in maintaining domain invariance.

Training procedure

The data were randomly divided into training and test datasets, resulting in an approximately 80:20 split. In the initial data splitting process, the dataset is divided into training and test sets according to the configured settings. The training set is used for model training, while the test set is used for final evaluation. During this process, the dataset is shuffled to ensure randomness. In the merging and filtering process, cognitive status and age data are merged and filtered according to specified criteria. This ensures that the training and test sets are balanced and representative of the overall dataset. Additionally, the training set is further divided into training and validation subsets to adjust model hyperparameters and evaluate performance during training. This process ensures that data does not overlap between the training and validation subsets, preventing data leakage and ensuring independent evaluation. In cases comprising multiple voice data from the same participant, we ensured that the same participant was not included in both sets.

Initially, CI and center classifier models were trained independently on a training set without domain generalization to establish their baseline performances (Fig. 1, dot line). Subsequently, domain generalization was applied, and the models were retrained. The cognition classifier utilizes the extracted content as input to produce predictions regarding CI. In tandem, the center classifier engages with the reconstructed output to predict center labels. During this phase, the models were optimized to minimize the influence of center-specific features on their predictions (Fig. 1, line).

To implement domain generalization, we introduced a transformation into the input data representation. This transformation was designed to ensure that the input features from different centers were indistinguishable, thereby minimizing the model's ability to learn center-specific characteristics. The feature extractor is trained with a gradient reversal layer that maximizes the center classifier's prediction error, meaning the model is trained to make it increasingly difficult for the domain classifier to accurately predict domains, thus enforcing a robust and generalized learning approach. Essentially, while classifiers for cognitive impairment and centers are concurrently trained, the optimization stepped up through

Table 3 Cognitive impairment and center classification results before and after applying domain generalization

DG	Classification	Accuracy	Precision	Recall	F1-score
Before	CI vs. NC	0.96	0.96	0.96	0.96
	Center	0.96	0.96	0.95	0.95
After	CI vs. NC	0.9	0.9	0.9	0.9
	Center	0.64	0.31	0.42	0.36

DG, domain generalization; CI, cognitive impairment; NC, normal cognition

Table 4 Confusion matrix for the center classification model using the validation dataset ($n = 58$)

		Predicted value			
		Center A	Center C	Center E	Center S
True value	Center A ($n = 29$)	26	0	0	3
	Center C ($n = 3$)	0	0	0	3
	Center E ($n = 12$)	10	0	0	2
	Center S ($n = 14$)	3	0	0	11

adversarial training enhances robustness against domain variance.

Evaluation metrics

After training, the models were validated on a separate validation set to ensure that the performance metrics were robust and reflected of the generalizability of the models. The validation involved calculating the precision, recall, and F1-scores for both CI and center classifications.

A confusion matrix was used to further analyze the performance of the center classification model. This analysis provided detailed insights into the misclassification rates for each center within the validation set. The decline in the post-domain-generalization performance on these metrics further substantiated the reduced ability of the models to distinguish between center-specific features.

Additionally, receiver operating characteristic (ROC) curves were plotted to compare the center classification performance before and after domain generalization. The area under the ROC curve (AUC) elucidated the model's ability to distinguish between the centers. The curves and corresponding AUC values revealed a significant decrease in the ability of the model to recognize center-specific characteristics post domain generalization.

Results

Before applying domain generalization, the CI classification model achieved an accuracy of 0.96. Additionally, when the center classification model was trained independently, it also achieved an accuracy of 0.96. Domain generalization was applied to transform the input data such that the centers could not be distinguished, with the aim to mitigate the unintentional learning of center characteristics by the CI classification model. After retraining

with domain generalization, the CI and center classification accuracies decreased from 0.96 to 0.90 and 0.96 to 0.64, respectively.

Table 3 shows the CI and center classification prediction results before and after applying domain generalization. The precision, recall, and F1-score represent the macro-averages of the results for each class. The confusion matrix of the center classification model after the domain generalization is presented in Table 4. The validation set included 29 individuals from Center A, 3 from Center C, 12 from Center E, and 14 from Center S. Of the 29 individuals from Center A, 3 were incorrectly predicted, whereas all 3 from Center C, 12 from Center E, and 3 out of 14 from Center S were misclassified.

Figure 2 presents the ROC curves for center classification with and without applying domain generalization; the left figure displays the ROC curves before ($AUC = 0.9944\text{--}1.000$) and the right figure shows those after domain generalization. Although the overall AUC decreased after domain generalization, Centers A (0.8210) and S (0.8815) exhibited relatively high AUC scores. However, those for Centers E (0.4755) and C (0.3303) decreased significantly, indicating a diminished domain effect.

Discussion

In this study, our primary objective was to reduce the impact of center-specific characteristics that could lead to biased predictions, focusing instead on generalizing cognitive features across different domains. Unlike other studies [4, 11, 13, 36, 37] that generally apply machine learning to voice data holistically, we employ denoising autoencoders specifically to isolate cognitive impairment (content) features from domain-specific (style) noise. This approach helps us to clearly isolate important cognitive features from background noise in our data, making the model better at working with recordings from different environments and reducing the effects of differences between centers. Unlike past studies that used the Deep Domain-Adversarial Image Generation (DDAIG) framework for image data, we applied it here specifically for detecting cognitive impairment through voice data. In this adapted framework, we split voice data into “content” (showing cognitive impairment features) and “style” (showing characteristics specific to each center), allowing us to separate relevant clinical information from background noise more effectively.

Our feature extractor employs a dual-encoder structure—one encoder for content and another for style—unlike traditional single-stream encoders [22]. This dual-encoding approach supports the explicit separation of cognitive features from environmental factors. The decoder then reconstructs the input by combining content and style, maintaining fidelity to the original input

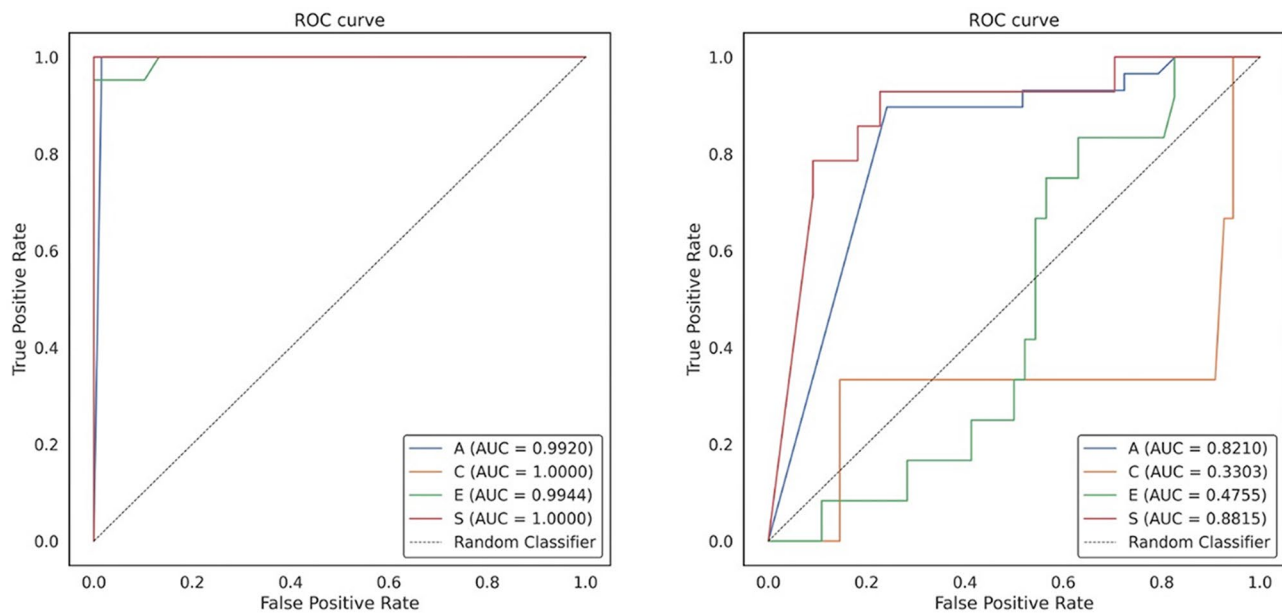


Fig. 2 ROC curves for center classification performance without (left) and with (right) domain generalization. The left panel shows high AUC scores for each center, indicating strong reliance on center-specific characteristics. In the right panel, the AUC scores decrease significantly after applying domain generalization, illustrating reduced center-specific influence

while effectively isolating clinically relevant features. This structural choice contrasts with previous implementations where content is treated as a slight modification of the original input.

The results, as shown in Table 3, indicate that while this approach led to a reduction in overall performance metrics such as accuracy, precision, recall, and F1-score, it succeeded in enhancing the model's ability to generalize. This shift suggests that the model prioritized cognitive features over center identification. We employed a DDAIG framework to ensure that the model learned cognitive features invariant across different centers, effectively mitigating the noise introduced by specific environmental conditions. Although this approach led to a decline in center classification accuracy, dropping from 0.96 to 0.64, it ultimately improved the model's generalizability, which is crucial for real-world applicability. The use of denoising autoencoders further helped disentangle “content” (cognitive impairment, CI) from “style” (center-specific characteristics), thereby reducing the risk of the model learning to associate irrelevant noise with cognitive features.

Despite the reduction in CI classification accuracy from 0.96 to 0.90 after implementing domain generalization, the model demonstrated a notable resilience. The relatively smaller decline in CI classification compared to center classification suggests that the core features critical for diagnosing cognitive impairment retained some robustness, even as center-specific influences were minimized.

The observed decline in the accuracy, precision, recall, and F1-score of the center classification model following the implementation of domain generalization raises important considerations. Initially, the model leveraged domain-specific features to enhance the accuracy of cognitive impairment (CI) classification. For example, domain characteristics such as the loudness of sound or the presence of echoes, which varied depending on the recording environment, inadvertently aided the model's performance. As a result, both the CI classification and center classification models achieved high accuracies of 0.96.

However, once domain generalization was applied, the model's focus shifted away from these domain-specific features, leading to a dilution of the information previously used to boost classification accuracy. The accuracy for the CI classification model decreased slightly to 0.90, while the center classification model experienced a more pronounced drop to 0.64. This suggests a trade-off: the model's ability to generalize across different datasets improved, but at the cost of its capacity to identify specific center features.

This trade-off highlights a critical challenge—the potential loss of contextual information that the model originally used to identify CI effectively. While the reduction in CI classification accuracy was less severe than in the center classification model, it is important to consider that the domain generalization process may have inadvertently removed some relevant features that contributed to accurate CI diagnosis. This outcome necessitates a closer examination of whether the broader changes

introduced by domain generalization have affected the feature representations crucial for identifying CI.

To address these issues, future work will focus on refining the training protocols to improve the model's sensitivity to various manifestations of CI while maintaining its generalization capabilities. Additional validation studies will be essential to ensure that the model continues to perform well across different domains without introducing new biases. Furthermore, expanding the dataset to include a more diverse sample of CI cases will be critical for enhancing the model's robustness and generalizability.

Although we tried to collect data that distinguished between patients with MCI and ADD, our goal was to identify individuals exhibiting early signs of cognitive decline. Therefore, we grouped MCI and ADD into a single class, referred to as "CI." Hence, the classification goal was to differentiate between NC and CI.

To prevent the CI classification model from inadvertently learning features specific to different centers, we implemented a domain generalization technique on the input data using a modified version of the DDAIG framework (Fig. 3) [33]. Figure 3 presents the architecture of our proposed model, designed to improve domain generalization by disentangling cognitive impairment-related features (content) from environment-specific characteristics (style). This model employs an autoencoder framework with two specialized encoders: the Content Encoder captures domain-invariant cognitive features, while the Style Encoder isolates information specific to the recording environment. The encoded content and style representations are combined in a Decoder to reconstruct the original input, validating that the encoders successfully separate cognitive and environmental features. The model's classification process includes two key components: (1) The Label Classifier, which uses content features to predict cognitive impairment, ensuring predictions are based on clinically relevant information. (2) The Domain Classifier, which classifies the

center-specific characteristics from the reconstructed input, highlighting any residual environmental information. In addition, to guide the model's learning, we used four loss functions: (1) Label loss, which optimizes cognitive impairment classification accuracy, (2) Domain loss, which focuses on center classification, (3) Negative Domain loss, applied to content features to minimize the presence of center-specific information, and (4) Autoencoder loss, which ensures the accurate reconstruction of the input.

Both the proposed and DDAIG models adhered to the architecture proposed in the study on domain-adversarial training of neural networks [20, 33]. Our research differs from Zhou et al. (2022) [20] and Farzana & Parde (2023) [38] in several ways. We utilize denoising autoencoders to effectively reconstruct clean cognitive features from noisy datasets, enhancing the reliability of our classifications. While previous studies treat features holistically, we distinctly separate content-related and style-related features to improve accuracy in cognitive impairment detection. Moreover, we have adapted the Deep Domain-Adversarial Image Generation (DDAIG) framework specifically for our task, allowing for better identification of cognitive impairments while minimizing domain-specific noise. The architecture of our study is composed of three fundamental components: feature extractor, label classifier, and domain classifier. Although the proposed and DDAIG models employed different structures for the feature extractor, both employed an identical function for transforming the input. In the model, these components are referred to as the feature extractor, cognition classifier, and center classifier, even though they perform equivalent functions.

The primary distinction lies in the use of a feature extractor. Although both models assign the same role to the feature extractor, the structural details differ slightly [20, 33]. In DDAIG, the feature extractor is called DoT-Net, which employs a series of residual blocks, followed

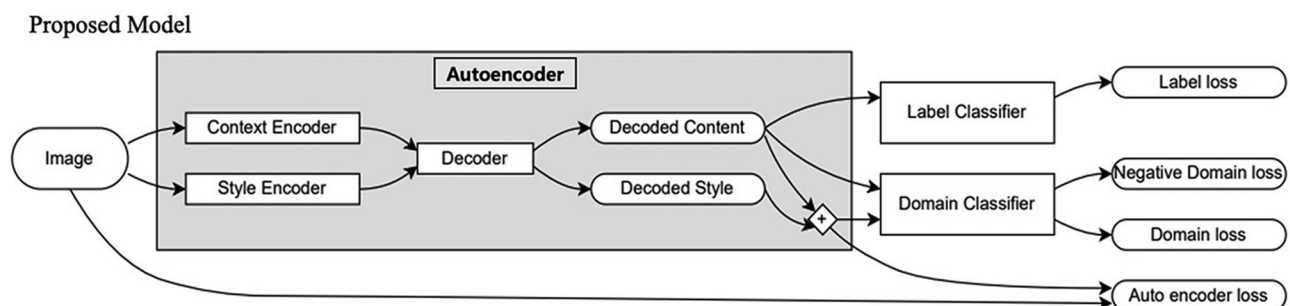


Fig. 3 Proposed domain generalization model. The model employs an autoencoder with a content encoder and a style encoder to decompose the input image into separate content and style representations. The content and style are decoded and then combined to reconstruct the input. The label classifier utilizes the decoded content to predict cognitive impairment labels, while the domain classifier uses the reconstructed input to identify center-specific features. The model is trained with multiple loss functions: label loss for cognitive classification accuracy, domain loss for center classification, negative domain loss to promote domain-invariance, and autoencoder loss to ensure accurate input reconstruction

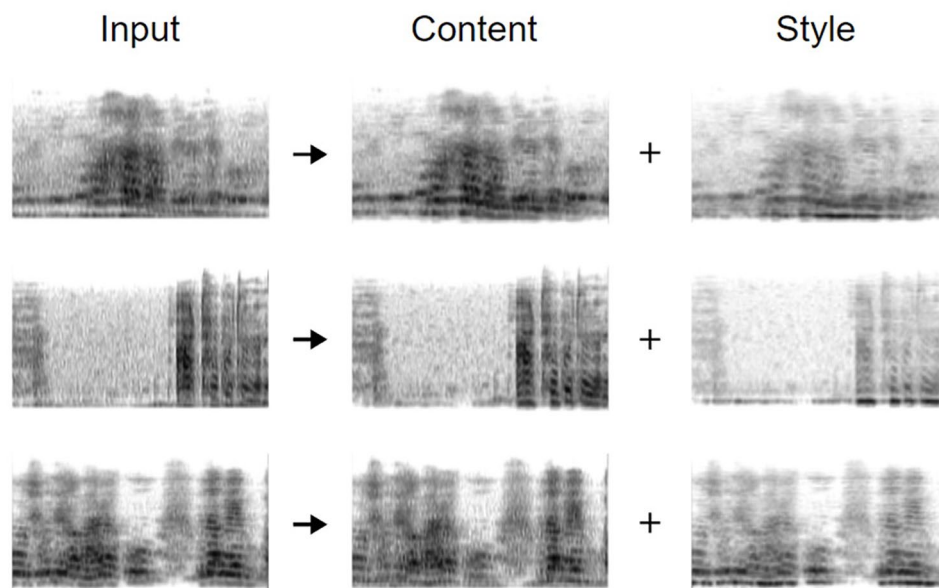


Fig. 4 Visual samples of the feature extractor inputs and outputs. Each row illustrates an example of the input mel-spectrogram (left), the extracted content representation (middle), and the extracted style representation (right). The content representation captures cognitive impairment features, while the style representation isolates environment-specific characteristics. When combined, these representations allow for the reconstruction of the input signal with a separation between cognitive-related and environment-specific features, enhancing the model's focus on domain-invariant cognitive indicators

by global average pooling that is concatenated with the output from these residual blocks. The result is a slightly transformed version of the original image, which we considered the content. Consequently, the content is equal to the input plus a modification (α), and DoTNet is trained to generate this modification.

The feature extractor of the proposed model comprises two encoders (content and style) and one decoder. The outputs from the two encoders are fed into the same decoder, resulting in two outputs representing content and style. This process allows content and style to be presented in any form unlike that in DDAIG, where content represents a modification. Therefore, we imposed constraints to ensure that the reconstructed output matched the input. Thus, the content plus style did not deviate from the original input. In other words, if the input (equal to the reconstructed output) equals content plus style, the content can be derived by deducting style from the input. Hence, the proposed model posits that by removing features related to the center (style), the content necessary for classifying cognitive disorders is retained.

Although the proposed model generated content and style, the content is used to classify labels and is intentionally not used for domain classification. Therefore, in the context of our model, the content and DoTNet outputs were equivalent.

Figure 2 presents the ROC curves for center classification with and without domain generalization, confirming that the center classification performance is affected by domain generalization. However, in the remaining

center groups (C, E, and S), where center S is the majority class, it can have a dominant influence. Indeed, by examining the confusion matrix of the center classification model, dominant predictions for Centers A and S can be observed along with a high recall for Center A. The overall AUC decreased after domain generalization, and Centers A and S still exhibited relatively high AUC values. However, the AUC scores for Centers E and C decreased significantly, indicating a diminished domain effect. Dominant predictions for Centers A and S, with a higher recall for Center A, suggest that the model favors the selection of Centers A and S, which have a relatively larger number of patients even after domain generalization.

The interpretation of these results indicated that the feature extractor played an important role in transforming the input data, and the CI classification model could distinguish between CIs. However, it faced difficulties during center classification, suggesting that the transformation by the feature extractor may have prioritized cognitive impairment classification over center classification.

As mentioned earlier, we hypothesized that practical transformations, such as noise removal, are necessary for effective classification. To confirm this, we investigated various statistics such as the mean, maximum, minimum, and median values for the input and output of the feature extractor, which revealed intriguing patterns. Figure 4 illustrates examples of a feature extractor that transforms the input data into content and style. Visual observations indicated that the mean and maximum values of the

content consistently decrease by approximately 60%–70% compared to those of the original input. Similarly, style showed a decrease of approximately 30–40% in the mean and maximum values. Contrary to our expectation that noise components unique to each center, such as distinctive sounds, would be included in style (expecting that they would disappear in content), content and style exhibited remarkably similar shapes, albeit with different value ranges. Despite these unexpected results, the feature extractor achieved the goal of rendering the centers indistinguishable. However, further research is necessary to elucidate the underlying cause of this success.

To strengthen the performance of the framework, several strategies could be considered. ‘Incorporating a multi-task learning’ framework, where the model learns to classify cognitive impairments while simultaneously reducing center-specific features through auxiliary tasks, could enhance performance by sharing information across tasks [39]. This approach enables the model to focus on domain-invariant features while minimizing overfitting. Another promising strategy is the use of ‘adaptive weighting’ mechanisms, which dynamically adjust the balance between cognitive impairment classification and domain generalization losses during training [40]. This ensures that both tasks are optimized effectively without compromising accuracy or generalization. Additionally, combining the DDAIG framework with ‘feature disentanglement networks and introducing contrastive loss’ could explicitly separate domain-invariant features, such as cognitive biomarkers, from domain-specific features, like recording conditions [33, 41]. Contrastive loss enforces a clear distinction between these features, making the learned representations more robust and reliable. These strategies could enhance the current framework and offer deeper insights into the interaction between domain-specific and domain-invariant features. Although these techniques could not be implemented in the present study, they represent valuable opportunities for future research to address limitations in the current approach and provide more robust solutions to domain shift challenges in cognitive impairment classification models.

Limitations and future directions

The sample diversity may not fully represent the broader population experiencing cognitive impairments. A limited sample size may restrict the generalizability of the findings. The findings call for additional validation studies to explore how well the model performs in real-world settings, which may differ significantly from controlled research environments. To validate the real-world performance of our AI diagnostic tool, prospective clinical studies are essential. These studies should involve diverse patient populations and multiple clinical settings

to assess the tool’s effectiveness, reliability, and generalizability. Real-world validation will provide critical insights into the practical application of our model and its potential impact on patient care.

We also recognize the necessity of benchmarking our methodology against established techniques. Therefore, we intend to assess these comparisons more thoroughly in future research phases. It would enhance its applicability in clinical settings.

Enhancing the predictive power of our model can be achieved by incorporating additional features or modalities. For instance, combining voice data with other types of data such as speech contents, meaningless voice, intersection, neuroimaging, genetic information, or cognitive test results could provide a more comprehensive assessment of cognitive disorders. Multimodal models may capture a wider range of biomarkers and improve diagnostic accuracy.

Voice-based cognitive disorder prediction models hold significant promise for use in everyday settings, not just clinical environments. Future research could explore the deployment of these models through wearable devices or smartphone applications, enabling continuous monitoring and early detection of cognitive impairments. This approach could facilitate timely interventions and improve long-term cognitive health management.

Finally, as AI tools are increasingly integrated into healthcare, ensuring data privacy and scalability is crucial. Future work should focus on developing robust data anonymization techniques and secure data handling protocols to protect patient privacy.

Conclusions

In this study, we demonstrated the potential of leveraging domain generalization techniques to enhance the robustness and applicability of voice-based CI detection systems across diverse clinical environments. Our findings suggest that, by integrating advanced domain adaptation strategies, the impact of dataset variability can be significantly reduced and the generalizability of the model to new unseen domains can be improved. Thus, this study paves the way for the development of more universally applicable diagnostic tools in the field of cognitive disorders, offering a promising avenue for future explorations to ensure equitable and effective healthcare solutions for various patient populations. To validate the robustness of this approach, it is essential to test it on data from new domains. Through the validation of new domains, we aim to enhance the reliability and generalization performance of the proposed method in real-world scenarios.

Abbreviations

AI	Artificial intelligence
DDAIG	Adversarial Image Generation
MCI	Mild cognitive impairment

CI	Cognitive impairment
NC	Normal cognition
AD	Alzheimer's disease ()
K-IADL	Korean version of Instrumental Activities of Daily Living Scale
KDSQ	Korean Dementia Screening Questionnaire
ADD	Alzheimer's Disease dementia
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve

Acknowledgements

We would like to thank Editage (www.editage.co.kr) for English language editing. We also would like to express our sincere gratitude to the patients and healthy elderly individuals from various hospitals and communities for their support in facilitating this study. This paper was written by the corresponding author, Youn YC, during sabbatical period. So, I extend my sincere gratitude to the members of the Department of Neurology at Chung-Ang University Hospital for their valuable support.

Author contributions

Y. C. Y. was responsible for study conception and design. S.K., Y.S, N.R., H.T.J., Y.C.Y., and H.C. participated in data collection. M.K., G.Y., W.Y. and H.L. contributed to statistical analysis of the data. M.K. and Y.C.Y. prepared the manuscript. All authors have reviewed the manuscript.

Funding

This research was supported by grants from the Ministry of SMEs and Startups (Project Number: S3079103), the Ministry of Education of the Republic of Korea, and the National Research Foundation of Korea (Project Number: NRF-2017S1A6A3A01078538). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. The funding organizations provided only financial support to enable the authors to conduct this research.

Data availability

The datasets generated and analyzed during this study are not publicly available in accordance with local ethics approval. However, they can be made accessible by the corresponding author in consultation with the institutional review board of the clinical centers (Seoul National University Bundang Hospital, Korean Nazarene University, Eunpyeong St. Mary's Hospital, and Chung-Ang University Hospital) on reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the IRBs of the respective clinical centers (Seoul National University Bundang Hospital IRB No, B-2109-709-303; Anseong City, IRB-Korean Nazarene University-2021-16; Eunpyeong St. Mary's Hospital, PC21ESDE0126; Chung-Ang University Hospital, 2110-028-481).

Informed consent

was obtained from all patients and/or their legal guardians, who signed the consent forms.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 February 2024 / Accepted: 28 October 2025

Published online: 24 November 2025

References

- Mahon E, Lachman ME. Voice biomarkers as indicators of cognitive changes in middle and later adulthood. *Neurobiol Aging*. 2022;119:22–35.
- Thomas JA, Burkhardt HA, Chaudhry S, Ngo AD, Sharma S, Zhang L, Au R, Hosseini Ghomi R. Assessing the utility of Language and voice biomarkers to predict cognitive impairment in the Framingham heart study cognitive aging cohort data. *J Alzheimers Dis*. 2020;76(3):905–22.
- Higuchi M, Nakamura M, Omiya Y, Tokuno S. Discrimination of mild cognitive impairment based on involuntary changes caused in voice elements. *Front Neurol*. 2023;14:1197840.
- Park CY, Kim M, Shim Y, Ryoo N, Choi H, Jeong HT, Yun G, Lee H, Kim H, Kim S, et al. Harnessing the power of voice: A deep neural network model for alzheimer's disease detection. *Dement Neurocogn Disord*. 2024;23(1):1–10.
- Khoshhal Mollasaraei Z, Behroozmand R. Impairment of the internal forward model and feedback mechanisms for vocal sensorimotor control in post-stroke aphasia: evidence from directional responses to altered auditory feedback. *Exp Brain Res*. 2024;242(1):225–39.
- Perry SE, Troche M, Huber JE, Curtis J, Kiefer B, Sevitz J, Dennard Q, Borders J, Browy JR, Dakin A, et al. Behavioral management of Respiratory/Phonatory dysfunction for dysarthria associated with neurodegenerative disease: A systematic review. *Am J Speech Lang Pathol*. 2024:1–29.
- Houde JF, Nagarajan SS, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: an MEG study. *J Cogn Neurosci*. 2002;14(8):1125–38.
- Houde JF, Jordan MI. Sensorimotor adaptation of speech I: compensation and adaptation. *J Speech Lang Hear Res*. 2002;45(2):295–310.
- Themistocleous C, Eckerstrom M, Kokkinakis D. Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *PLoS ONE*. 2020;15(7):e0236009.
- König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, Manera V, Verhey F, Aalten P, Robert PH, et al. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimers Dement (Amst)*. 2015;1(1):112–24.
- Garcia-Gutierrez F, Marquie M, Munoz N, Alegret M, Cano A, de Rojas I, Garcia-Gonzalez P, Olive C, Puerta R, Orellana A, et al. Harnessing acoustic speech parameters to Decipher amyloid status in individuals with mild cognitive impairment. *Front Neurosci*. 2023;17:1221401.
- Taptiklis N, Su M, Barnett JH, Skirrow C, Kroll J, Cormack F. Prediction of mental effort derived from an automated vocal biomarker using machine learning in a large-scale remote sample. *Front Artif Intell*. 2023;6:1171652.
- Parsapoor Parsa MM, Alam MR, Mihailidis A. Performance of machine learning algorithms for dementia assessment: impacts of Language tasks, recording media, and modalities. *BMC Med Inf Decis Mak*. 2023;23(1):45.
- Tigga NP, Garg S, Goyal N, Raj J, Das B. Brain-region specific autism prediction from electroencephalogram signals using graph Convolution neural network. *Technol Health Care*. 2025;33(1):77–101.
- Garg NPTS, Alnajjar F. Autism Spectrum Disorder Classification with EEG Signals Using Dense Graph. In: *Converging Clinical and Engineering Research on Neurorehabilitation V: Proceedings of the 6th International Conference on Neurorehabilitation (ICNR 2024)*, November 5–8, 2024, La Granja, Spain-2: 2024: Springer Nature; 2024: 350.
- Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet Classifiers Generalize to ImageNet? In: *Proceedings of the 36th International Conference on Machine Learning*. Edited by Kamalika C, Ruslan S, 97 Proceedings of Machine Learning Research: PMLR; 2019: 5389–5400.
- Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. *Mach Learn*. 2010;79(1):151–75.
- Hires M, Drotar P, Pah ND, Ngo QC, Kumar DK. On the inter-dataset generalization of machine learning approaches to parkinson's disease detection from voice. *Int J Med Inf*. 2023;179:105237.
- Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med*. 2018;15(11):e1002683.
- Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: A survey. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(4):4396–415.
- Lee WH, Ozger M, Challita U, Sung KW. Noise Learning-Based denoising autoencoder. *IEEE Commun Lett*. 2021;25(9):2983–7.
- Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th international conference on Machine learning*: 2008; 2008: 1096–1103.
- Zhang Y, Zhang Y, Cai W. A unified framework for generalizable style transfer: style and content separation. *IEEE Trans Image Process*. 2020.
- Mercy Faustina J, Akash V, Gupta A, Divya V, Manoj T, Sadagopan N, Sivaselvan B. A study of neural Artistic style transfer models and architectures for Indian Art styles. *Network*. 2023;34(4):282–305.
- Gatys LA, Ecker AS, Bethge M. A Neural Algorithm of Artistic Style. *arXiv 2015: eprint arXiv:1508.06576*.
- Dixon JR Jr. The International Conference on Harmonization Good Clinical Practice guideline. *Qual Assur* 1998, 6(2):65–74.

27. Christensen KJ, Multhaup KS, Nordstrom S, Voss K. A cognitive battery for dementia: development and measurement characteristics. *Psychol Assessment: J Consulting Clin Psychol*. 1991;3(2):168–74.
28. Kang SC, Lee S, Kwon B, Na J, Han DL. The reliability and validity of the Korean instrumental activities of daily living (K-IADL). *J Korean Neurol Assoc*. 2002;20(1):8–14.
29. Kim A, Kim S, Park KW, Park KH, Youn YC, Lee DW, Lee JY, Lee JH, Jeong JH, Choi SH, et al. Reliability and validity of a short form of the Korean dementia screening Questionnaire-Cognition. *J Clin Neurol*. 2020;16(1):145–53.
30. Landau HJ. Sampling, data transmission, and the Nyquist rate. *Proc IEEE*. 1967;55(10):1701–6.
31. Ariyanti W, Liu KC, Chen KY, Yu T. Abnormal respiratory sound identification using Audio-Spectrogram vision transformer. *Annu Int Conf IEEE Eng Med Biol Soc*. 2023;2023:1–4.
32. Gong Y, Chung Y, Glass J. AST: Audio Spectrogram Transformer. *arXiv*, 2021. preprint arXiv:2104.01778.
33. Zhou K, Yang Y, Hospedales T, Xiang T. Deep Domain-Adversarial Image Generation for Domain Generalisation. *arXiv 2020:arXiv:2003.06054*.
34. Ganin YUE, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V. Domain-Adversarial training of neural networks. *J Mach Learn Res*. 2016;17:1–35.
35. Gallego AJ, Calvo-Zaragoza J, Fisher RB. Incremental unsupervised Domain-Adversarial training of neural networks. *IEEE Trans Neural Netw Learn Syst*. 2021;32(11):4864–78.
36. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham heart study. *Alzheimers Res Ther*. 2021;13(1):146.
37. Konig A, Satt A, Sorin A, Hoory R, Derreumaux A, David R, Robert PH. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Curr Alzheimer Res*. 2018;15(2):120–9.
38. Farzana SP, Natalie. Towards Domain-Agnostic and Domain-Adaptive Dementia Detection from Spoken Language. In: the 61st Annual Meeting of the Association for Computational Linguistics 2023; Toronto, Canada: Association for Computational Linguistics. 2023: 11965–11978.
39. Ruder S. An overview of Multi-Task learning in deep neural networks. *ArXiv*; 2017.
40. Gong T, Lee T, Stephenson C, Renduchintala V, Padhy S, Ndirango A, Keskin G, Elibol OH. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*. 2019;7:141627–32.
41. Guan Y, Han J, Song H, Song W, Zheng G, Zheng T, He Y. Contrastive Loss Based Frame-wise Feature disentanglement for Polyphonic Sound Event Detection. *ArXiv* 2024.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.