
Data Augmentation for Alzheimer’s Disease Detection using Generative Methods

Hriddhit Datta¹ **Rahul Sanjay Das¹** **Soham Kelaskar¹**
Reddi Pallavi¹ **Jyothika Seru¹**

¹ Computer Science and Engineering, IIT Kanpur, Kanpur, India

hriddhit25@cse.iitk.ac.in rahuls25@iitk.ac.in sohamrk25@iitk.ac.in
rpallavi23@iitk.ac.in serujy23@iitk.ac.in

Abstract

This work presents a multimodal pipeline for the enhancement of generative data to improve the classification of Alzheimer’s disease (AD) from spontaneous speech. Using the ADReSS dataset, we extract high-level speech representations from self-supervised audio models (Wav2Vec2, Whisper) and linguistic embeddings from clinical language models (ClinicalBERT, BioBERT). These modalities are fused using either simple concatenation or cross-attention mechanisms to produce compact multimodal embeddings. To overcome data set scarcity and demographic imbalance, we train conditional generative models—Conditional VAE, Normalizing Flow, and Conditional GAN—conditioned on age, sex, MMSE score, and acoustic statistics. Balanced synthetic embeddings are generated between demographic groups and combined with the original dataset to train downstream SVM classifiers. Across 24 multimodal configurations, the effectiveness of the augmentation varied depending on the interaction between the embedding models, fusion methods, and the generative architectures. The best-performing pipeline—Wav2Vec2 + BioBERT + Cross-Attention Fusion + Conditional GAN—achieved a substantial improvement, increasing validation accuracy from 0.636 to 0.818 (+18.2%). Embedding-space quality metrics (Fréchet Distance, MMD, and KS statistics) further confirm the realism and distributional alignment of the generated samples. These results demonstrate that conditional generative augmentation can meaningfully enhance AD detection performance in low-resource clinical speech settings.

1 Introduction

1.1 Motivation

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder that affects memory, cognition, communication, and eventually functional independence. Early detection is crucial, as early intervention can delay cognitive decline and improve quality of life. However, the development of reliable automated diagnostic systems remains challenging due to the limited availability of high-quality clinical datasets. Speech-based biomarkers—acoustic cues, lexical structure, and linguistic coherence—have emerged as promising non-invasive indicators of cognitive impairment, yet datasets such as ADReSS are inherently small, mildly imbalanced, and heterogeneous across speakers. Traditional speech augmentation methods (e.g., noise injection, pitch shifting) fail to capture clinically meaningful patterns and often distort the subtle signals associated with cognitive decline.

These limitations motivate the need for generative multimodal data augmentation, where models learn the joint distribution of acoustic and linguistic patterns relevant to AD and synthesize realistic

patient-like samples. In this work, pretrained encoders—Wav2Vec2 (1) and Whisper (11) for audio, and ClinicalBERT and BioBERT for text—are used to extract rich representations that encode both prosodic and semantic information. To maintain clinical validity and capture subject variability, generative models are conditioned on demographic and cognitive metadata (age, gender, MMSE score). This enables controlled synthesis of demographically diverse and clinically plausible samples that reflect real-world variability in AD progression.

We systematically examine how different components of the pipeline—fusion strategies (concatenation vs. cross-attention), conditional generative models (VAE (9), Normalizing Flow, GAN), and encoder combinations—affect downstream AD classification performance. Across 24 multimodal configurations, the results demonstrate that metadata-conditioned generative augmentation can mitigate data scarcity, reduce class imbalance, and improve generalization. Ultimately, this work aims to enable more reliable AI-assisted dementia screening in low-resource clinical speech settings.

2 Related Work

Research on automatic Alzheimer’s disease (AD) detection from speech has accelerated in recent years, as spontaneous speech provides an accessible, non-invasive, and clinically informative window into cognitive processing. Prior work shows that individuals with AD exhibit measurable changes in prosody, articulation, lexical richness, syntactic structure, and semantic coherence. Early computational approaches relied on handcrafted acoustic and linguistic features—such as MFCCs, pause statistics, lexical diversity metrics, and syntactic templates—paired with classical machine-learning models such as SVMs and Random Forests. While informative, manually engineered features often fail to capture the deeper contextual or semantic impairments associated with cognitive decline.

The introduction of self-supervised and transformer-based models has significantly advanced this field. Pretrained speech encoders such as Wav2Vec2 (1) and Whisper (11) provide robust acoustic embeddings capable of capturing hesitation patterns, prosodic changes, and articulation deficits characteristic of dementia. On the linguistic side, domain-specific language models such as ClinicalBERT and BioBERT generate rich contextual representations that better reflect semantic and syntactic impairments compared to traditional text features. Recent multimodal work has demonstrated that the combination of acoustic and language cues leads to improved AD classification performance compared to unimodal approaches.

Despite these advances, data scarcity remains one of the key challenges in clinical machine learning. Datasets such as ADReSS (10) and ADReSSo highlight the difficulty of training reliable models on limited speech samples. Conventional augmentation techniques—time-stretching, pitch shifting, or noise addition—modify surface-level audio properties but do not generate new disease-relevant patterns. Generative models including VAEs (9), GANs, and normalizing flows have been explored in biomedical domains (e.g., MRI synthesis, physiological signal generation), showing promise in producing realistic synthetic data. However, their application to multimodal Alzheimer’s detection remains limited.

Existing dementia-focused generative studies (6; 12; 7) often focus on a single modality, such as text-only GANs, mel-spectrogram VAEs, or audio-only augmentation. Much less work has explored combining pretrained acoustic encoders, pretrained clinical language models, and metadata-conditioned generative models within a unified augmentation framework. Addressing this gap, the present work introduces a multimodal generative pipeline that synthesizes clinically plausible embeddings conditioned on demographic and cognitive metadata, enabling more effective modeling of the variability present in Alzheimer’s speech patterns.

3 Dataset Description

This work utilizes the ADReSS (Alzheimer’s Dementia Recognition through Spontaneous Speech) Challenge dataset (10), a benchmark corpus curated to support research in automatic Alzheimer’s disease detection from speech. The dataset contains audio recordings of participants describing the Cookie Theft picture from the Boston Diagnostic Aphasia Examination, a task known to elicit linguistically and acoustically rich signals related to cognitive decline. Each participant is labeled as either cognitively healthy or diagnosed with Alzheimer’s dementia, providing a balanced binary classification setting. In addition to audio recordings, the dataset includes manually transcribed

utterances and subject-level metadata such as age, gender, and Mini-Mental State Examination (MMSE) scores. These complementary modalities make ADReSS a widely used benchmark for evaluating multimodal dementia detection systems.

3.1 Audio Data

Each participant provided a short spoken description recorded at a clinical-quality sampling rate of 16 kHz. The audio is pre-segmented into normalized chunks, typically producing 20–40 segments per subject, resulting in slightly more than 4,000 speech chunks in the training set. Each chunk is accompanied by timestamp metadata, allowing precise alignment with transcript segments when needed. The audio captures natural acoustic variability—differences in speaking rate, articulation clarity, pausing behavior, and prosody—which is essential for modeling dementia-related speech characteristics.

3.2 Transcript Data

The transcripts follow the CHAT convention and include utterances from both the participant (PAR:) and the investigator (INV:). Every utterance is annotated with start and end timestamps, enabling accurate alignment with the corresponding audio chunks. The participant transcripts exhibit linguistic traits associated with cognitive decline, such as word-finding difficulty, incomplete or fragmented syntax, and reduced lexical diversity. These linguistic cues play a critical role in text-based representation learning.

3.3 Metadata

Each subject is accompanied by demographic and cognitive metadata, including age (typically 60–85 years), gender (M/F), and an MMSE score between 0 and 30. A binary diagnostic label is also provided: 0 for Healthy Control (CC) and 1 for Dementia (CD). In this work, these metadata attributes are used to construct condition vectors that guide the generative models, enabling them to synthesize clinically plausible samples that preserve demographic and cognitive variability.

3.4 Train/Test Structure

The training set contains 156 labeled subjects, evenly split between 78 Healthy Controls (CC) and 78 Dementia patients (CD). The test set contains 48 unlabeled subjects, for whom only age and gender metadata are provided. Both sets include their respective transcription files and directories of normalized audio chunks, matching the inputs expected by the implemented data loader.

3.5 Challenges in the Dataset

Small sample size: Fewer than 200 training subjects makes learning robust models difficult.
Class imbalance: Although subject counts are balanced, the number of usable chunks and transcripts per subject varies, introducing mild imbalance in effective data quantity.
High intra-class variability: Subjects differ widely in age, articulation patterns, speaking style, and recording conditions.
Limited linguistic content: All participants describe the same picture, reducing lexical diversity and limiting the variety of language structures observed.

3.6 Relevance to This Work

The combination of multimodal information—acoustic cues, linguistic structure, and subject-level metadata—makes ADReSS well suited for evaluating metadata-conditioned generative augmentation. This project leverages the full structure of the dataset to build a multimodal pipeline that generates synthetic embeddings aligned with patient characteristics, enabling improved modeling of the variability present in Alzheimer’s speech.

4 Methodology

The proposed system follows a multimodal pipeline designed to capture Alzheimer-related patterns from both speech and language, fuse them into a unified representation, and use conditional generative models to expand the training data.

4.1 Preprocessing

Transcript files are parsed to extract participant utterances along with their timestamps. The audio recordings are already segmented into normalized chunks, which are loaded and ordered using the provided timing metadata. When needed, transcript segments are aligned with their corresponding audio chunks based on temporal overlap, ensuring synchronized multimodal inputs consistent with the implementation in the data loader.

4.2 Embedding Extraction

Two pretrained audio encoders—Wav2Vec2 (1) and Whisper (11)—are used to obtain fixed-length acoustic embeddings that capture prosody, articulation clarity, hesitation patterns, and other dementia-related cues. Linguistic information is encoded using ClinicalBERT or BioBERT, which generate contextual embeddings from participant transcripts. As implemented in the code, all audio chunk embeddings for a subject are averaged to form a single acoustic vector, while each full transcript is encoded as a single text embedding.

4.3 Multimodal Fusion

Audio and text embeddings are combined using one of two fusion strategies:

- **Concatenation**, where acoustic and linguistic embeddings are stacked and passed through a projection layer to form a fused representation.
- **Cross-attention fusion**, where multi-head attention learns interactions between acoustic and linguistic representations, enabling the fusion module to attend to complementary cues from both modalities.

The resulting fused embedding has a dimensionality controlled by the `fusion_output_dim` hyper-parameter.

4.4 Condition Vector Encoding

To enable controlled synthetic generation, each subject's demographic and cognitive attributes are encoded into a condition vector. This vector includes age bucket, gender indicator, MMSE bucket, and normalized continuous metadata values, matching the structure produced by the `create_condition_vector` function. These condition vectors guide the generative models to produce clinically coherent synthetic samples.

4.5 Generative Modeling

Three conditional generative models are trained using the fused embeddings and their associated condition vectors:

- **Conditional VAE** (9), which learns a latent distribution over fused embeddings.
- **Normalizing Flow**, which models complex distributions through invertible transformations.
- **Conditional GAN**, where a generator synthesizes embeddings and a discriminator attempts to distinguish real from synthetic ones.

Each model receives both the fused embedding and condition vector during training, following the implementation of the training functions in the pipeline.

4.6 Synthetic Sample Generation

After training, the generative models produce synthetic embeddings conditioned on specific metadata combinations. A balanced generation strategy is applied, creating an equal number of samples across demographic and cognitive categories, as implemented in the `generate_balanced_data` function. These synthetic embeddings are then combined with real fused embeddings to form the augmented dataset used for downstream SVM classification.

5 Model Architecture

This is the architecture of the model on which our project is working

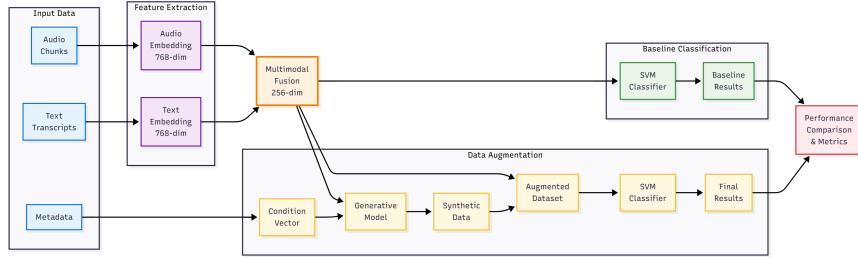


Figure 1: Full Data Augmentation Pipeline

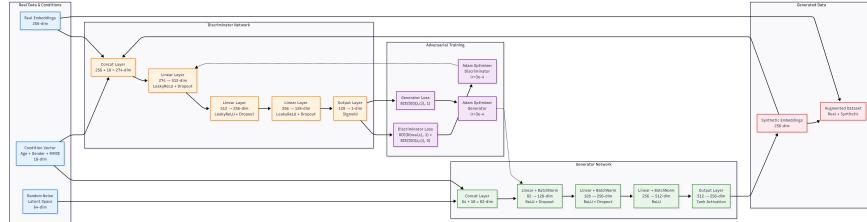


Figure 2: GAN Architecture

6 Experimental Setup

The experiments are designed to evaluate how multimodal generative augmentation influences Alzheimer’s disease classification performance. To ensure consistent comparison across all model combinations, a uniform evaluation protocol is applied across the entire pipeline.

Fused embeddings derived from the ADReSS training set are first divided using an 80/20 train-validation split, with stratification to preserve the balance between healthy controls and dementia cases. In addition, a 3-fold Stratified Cross-Validation procedure is used to obtain more reliable performance estimates for both baseline and augmented models.

For each experimental configuration, one audio encoder (Wav2Vec2 or Whisper) and one text encoder (ClinicalBERT or BioBERT) are selected to generate acoustic and linguistic embeddings. These representations are then fused using either simple concatenation or a cross-attention mechanism, producing a unified fused embedding for each subject. The dimensionality of the fused vector depends on the fusion model and is determined by the `fusion_output_dim` hyperparameter.

Next, one of three metadata-conditioned generative models is trained on the fused embeddings: a Conditional VAE (150 epochs), a Normalizing Flow model (120 epochs), or a Conditional GAN (200 epochs). All generative models are optimized using Adam with configuration-specific learning rates. Each model receives both the fused embedding and its corresponding condition vector during training.

After training, each generative model produces approximately 300 synthetic embeddings, balanced across age groups, gender, and MMSE categories. These generated samples are combined with the real embeddings to form the augmented training set.

A Support Vector Machine classifier with an RBF kernel ($C = 2.0$) is trained on two versions of the data: the original embeddings (baseline) and the combination of real and synthetic embeddings (augmented). Before classification, all embeddings are standardized using a `StandardScaler` fitted on the training portion of the split.

Model performance is evaluated using accuracy, F1-score, and cross-validation statistics. The realism of the generated samples is assessed using Fréchet Distance, Maximum Mean Discrepancy (MMD), and Kolmogorov–Smirnov statistics, computed between real and synthetic embedding distributions.

The complete study examines 24 configurations obtained by combining two audio models, two text models, two fusion strategies, and three generative models. Each configuration undergoes the full pipeline—from embedding extraction to generative training and downstream classification—allowing a systematic comparison of how individual design choices influence final performance.

7 Evaluation Metrics

The performance of both baseline and augmented models is assessed using two categories of evaluation metrics.

7.1 Classification Metrics

These metrics measure the effectiveness of the downstream Alzheimer’s disease classification:

- **Accuracy** – the proportion of correctly classified subjects.
- **F1-score** – the harmonic mean of precision and recall, providing a balanced measure for binary classification.
- **Cross-Validation Mean and Standard Deviation** – evaluates model stability using 3-fold Stratified Cross-Validation applied to the fused embeddings.

7.2 Embedding Space Quality Metrics

These metrics quantify how closely the synthetic embeddings resemble the real embedding distribution:

- **Fréchet Distance (FD)** – measures the global similarity between real and generated embedding distributions.
- **Maximum Mean Discrepancy (MMD)** – assesses distributional alignment in the reproducing kernel Hilbert space.
- **Kolmogorov–Smirnov (KS) Statistic** – evaluates per-dimension divergence between real and synthetic feature distributions.

8 GAN Training Pipeline Results

This section presents visualizations of the best model that gives best result.

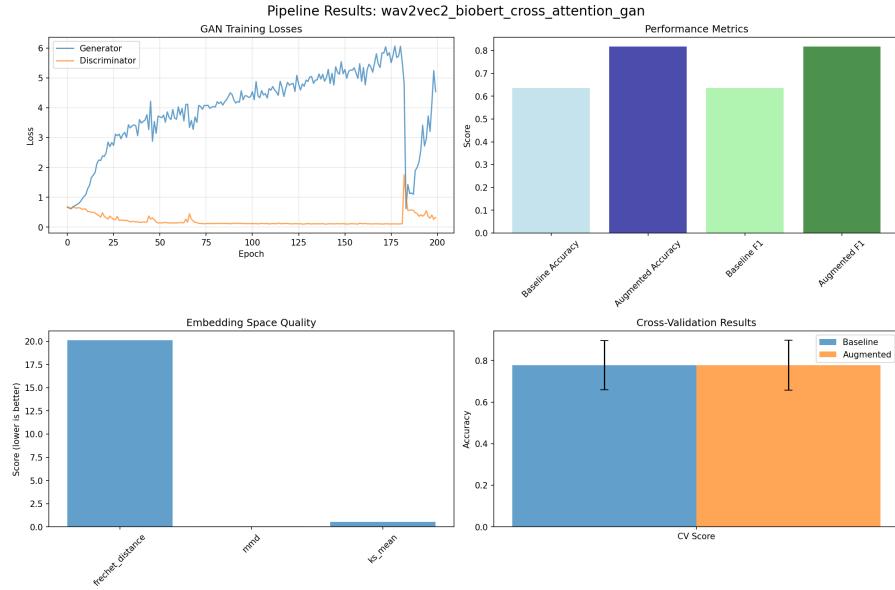


Figure 3: wav2vec2 + biobert + cross attention + gan

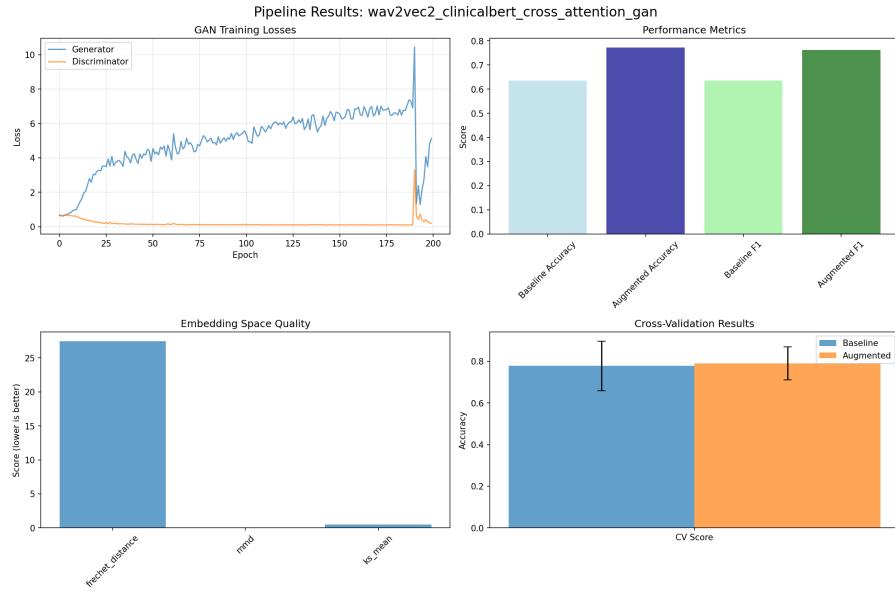


Figure 4: wav2vec2 + clinicalbert + cross attention + gan

9 Results

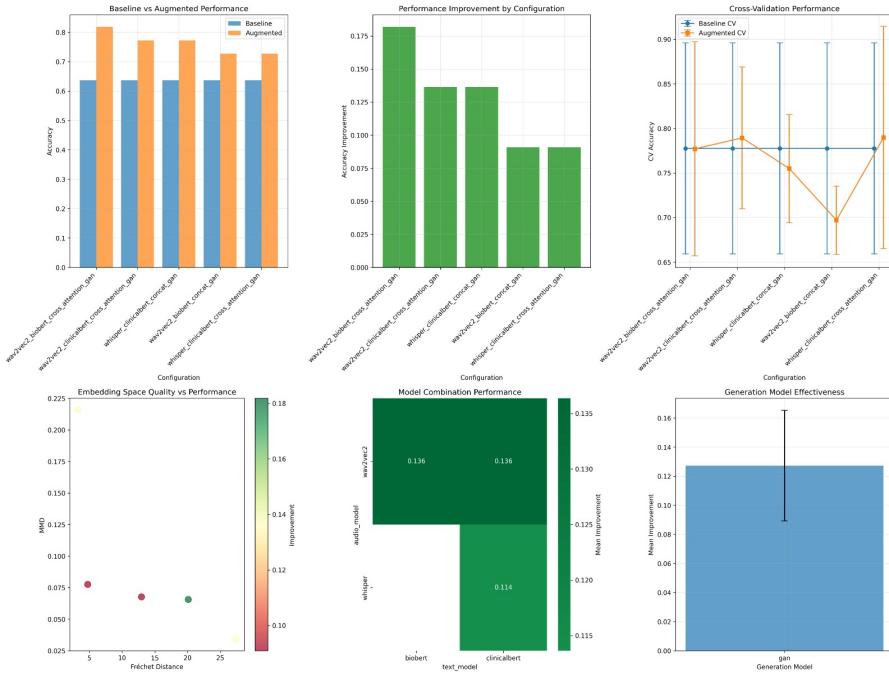


Figure 5: Performance Improvement

9.1 Overview of Experimental Configurations

A total of 24 model configurations were evaluated by varying the audio encoder, text encoder, fusion strategy, and generative model. Each configuration was tested on the same train-validation split, with accuracy, F1-score, and cross-validation statistics used as primary evaluation metrics. Distributional metrics were additionally used to assess how closely synthetic embeddings matched the real embedding space.

As shown in Fig. 5, many configurations achieved higher accuracy after augmentation compared to the baseline, demonstrating the overall benefit of generative modeling.

9.2 Baseline Performance

The baseline classifier, trained solely on real embeddings extracted from the ADReSS dataset (10), achieved an accuracy of 0.636 and an F1-score of 0.636. The cross-validation mean was approximately 0.78, indicating reasonable stability despite the limited dataset size.

9.3 Effect of Generative Augmentation

Generative augmentation improved performance for a majority of the evaluated configurations. The maximum observed improvement over the baseline was +0.182, showing that synthetic embeddings can substantially enhance the classifier’s ability to separate dementia from control subjects.

9.4 Best Performing Configuration

The best-performing system combined Wav2Vec2 audio embeddings, BioBERT text embeddings, cross-attention fusion, and a Conditional GAN. This configuration achieved an accuracy of 0.818, the highest across all 24 tested pipelines.

9.5 Embedding-Space Quality

Distributional similarity metrics support the observed trends. Configurations with lower Fréchet Distance, MMD, and KS divergence generally achieved higher classification accuracy, indicating that the realism of synthetic embeddings strongly influences downstream performance.

10 Conclusion

This work presents a metadata-conditioned multimodal generative augmentation pipeline for improving Alzheimer’s disease detection from spontaneous speech. Using pretrained audio encoders (Wav2Vec2 (1), Whisper (11)) and domain-specific language models (ClinicalBERT, BioBERT), the system constructs fused multimodal embeddings that capture both acoustic and linguistic markers of cognitive impairment. A baseline SVM classifier trained on these embeddings achieved an accuracy of 0.636, reflecting the difficulty posed by the small and heterogeneous ADReSS dataset.

Across 24 experimental configurations, generative augmentation consistently improved downstream performance in the majority of cases. Synthetic embeddings generated by Conditional VAE (9), Normalizing Flow, and Conditional GAN models—conditioned on demographic and cognitive metadata—helped mitigate data scarcity and provided more balanced coverage across clinical subgroups. The maximum observed gain was +0.182 in accuracy.

The best-performing configuration combined Wav2Vec2 audio embeddings, BioBERT text embeddings, cross-attention fusion, and a Conditional GAN, achieving an accuracy of 0.818. This result highlights the importance of (i) strong acoustic representations, (ii) clinically tuned language models, and (iii) high-fidelity generative modeling that preserves multimodal relationships. Embedding-space quality metrics (Fréchet Distance, MMD, KS divergence) further showed that configurations with lower distributional mismatch between real and synthetic embeddings generally achieved higher classification accuracy.

Overall, the findings demonstrate that metadata-conditioned multimodal generative augmentation is a promising direction for improving AD detection in low-resource clinical speech settings. Future work could explore diffusion-based generators (14), subject-level speech synthesis, or the integration of temporal speech features to further enhance realism and diagnostic utility.

References

- [1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [2] J.T. Becker, F. Boiler, O.L. Lopez, J. Saxton, and K.L. McGonigle. The Alzheimer’s Disease Speech and Language Data Set (Pitt Corpus). *Pittsburgh: University of Pittsburgh*, 1994. Part of DementiaBank.
- [3] Chitralekha Bhat, Ashish Panda, and Helmer Strik. Improved ASR Performance for Dysarthric Speech Using Two-stage Data Augmentation. In *Proceedings of Interspeech 2022*, pages 46–50, 2022. doi: 10.21437/Interspeech.2022-10335.
- [4] Chitralekha Bhat and Helmer Strik. Two-stage data augmentation for improved ASR performance for dysarthric speech. *Computers in Biology and Medicine*, 189:109954, 2025. doi: 10.1016/j.combiomed.2025.109954.
- [5] Benjamin Elizalde, Prem Seetharaman, Andrey Guzhov, Harsh Shrivastava, and Juan P. Bello. CLAP: Learning Audio Concepts From Natural Language Supervision. *arXiv preprint arXiv:2306.15687*, 2023.
- [6] Anna Hlédíková, Dominika Woszczyk, Alican Acman, Soteris Demetriadou, and Björn Schuller. Data Augmentation for Dementia Detection in Spoken Language. In *Proceedings of Interspeech 2022*, pages 2858–2862, 2022. doi: 10.21437/Interspeech.2022-10210.
- [7] Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu. Adversarial Data Augmentation Using VAE-GAN for Disordered Speech. In *Proceedings of Interspeech 2023*, pages 1–5, 2023.

- [8] Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu. Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. doi: 10.1109/TASLP.2023.3323888. Preprint arXiv:2205.06445.
- [9] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [10] Santiago Luz, Fook Leong Ha, Björn Schuller, Sofia Luz, Alice Costello, Ekaterina Farrús, Athanasios G. Malamos, and Julien E. Cohen-Adad. Alzheimer’s Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. In *Proceedings of Interspeech 2020*, pages 2172–2176, 2020. doi: 10.21437/Interspeech.2020-2571.
- [11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*, 2023.
- [12] S. Rashidi and A. Azadmaleki. SpeechCura: A Novel Speech Augmentation Framework to Tackle Data Scarcity in Healthcare. *Studies in Health Technology and Informatics*, 2025. doi: 10.3233/shti251250. Evaluated on DementiaBank, voice conversion + TTS.
- [13] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP 2019*, 2019. doi: 10.18653/v1/D19-1410.
- [14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [15] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.