



On the Within-class Variation Issue in Alzheimer’s Disease Detection

Jiawen Kang¹, Dongrui Han², Lingwei Meng¹, Jingyan Zhou¹, Jinchao Li¹, Xixin Wu¹,
Helen Meng^{1,2}

¹The Chinese University of Hong Kong, Hong Kong SAR, China

²Centre for Perceptual and Interactive Intelligence, Hong Kong SAR, China

jwkang@se.cuhk.edu.hk

Abstract

Alzheimer’s Disease (AD) detection employs machine learning classification models to distinguish between individuals with AD and those without. Different from conventional classification tasks, we identify within-class variation as a critical challenge in AD detection: individuals with AD exhibit a spectrum of cognitive impairments. Therefore, simplistic binary AD classification may overlook two crucial aspects: within-class heterogeneity and instance-level imbalance. In this work, we found using a sample score estimator can generate sample-specific soft scores aligning with cognitive scores. We subsequently propose two simple yet effective methods: Soft Target Distillation (SoTD) and Instance-level Re-balancing (InRe), targeting two problems respectively. Based on the ADReSS and CU-MARVEL corpora, we demonstrated and analyzed the advantages of the proposed approaches in detection performance. These findings provide insights for developing robust and reliable AD detection models.

Index Terms: Alzheimer’s disease, neurocognitive disorder, within-class variations, AD detection, dementia, healthcare

1. Introduction

Neurocognitive disorders (NCD) such as Alzheimer’s Disease (AD), present a substantial and growing challenge within the aging population, characterized by progressive cognitive decline across multiple domains, including memory, attention, and executive function [1]. For timely intervention and management, in-person clinical assessments have been the primary protocol for screening AD patients in populations, where participants are examined using specially designed assessment tasks to test potential abnormal declines in cognitive abilities [2–4]. In contrast to traditional on-set assessments, recent advancements in machine learning technologies have facilitated speech-based automatic AD detection as a promising screening approach, with the advantages of being scalable, accessible, and cost-efficient.

Common practices of machine learning AD detection are to model this task as binary classifications, i.e., prediction models are trained on audio recordings or transcripts from assessment tasks to classify participants as AD or non-AD [5–7]. This modeling largely inherits the paradigm of standard machine learning classification, which is dedicated to extracting discriminative features or patterns and then deploying a classifier for prediction. In recent years, the progress in AD detection has been largely driven by the exploration of effective features and representation learning. As a brief review, early works leveraged handcrafted acoustic and linguistic features. For example, Alhanai et al. [8] identified 12 acoustic features, including decreasing jitter, strongly associated with cognitive impairment.

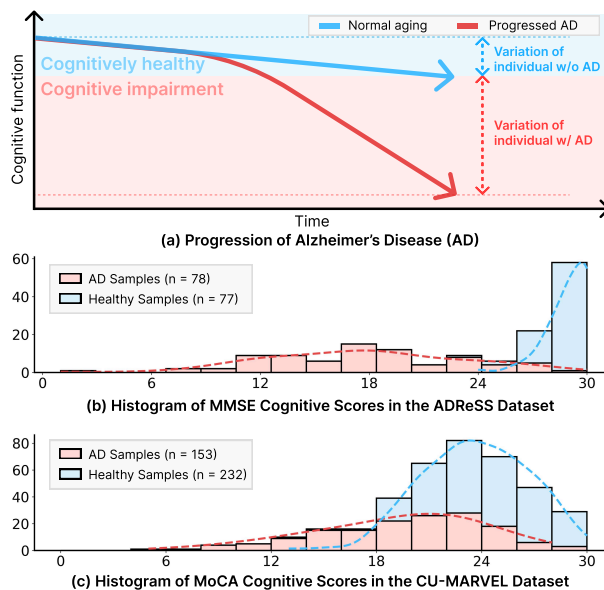
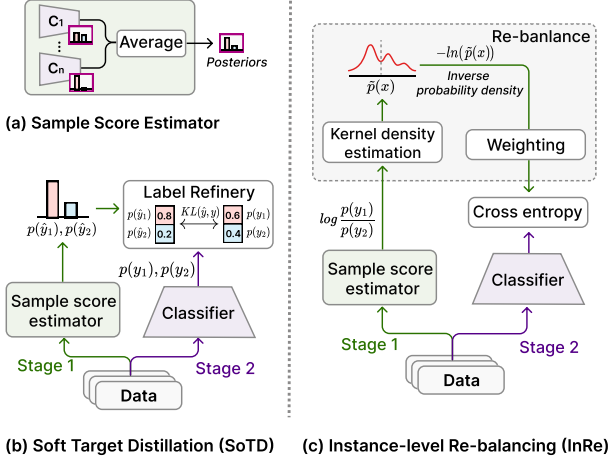


Figure 1: Visualizations of the within-class variation in Alzheimer’s disease (AD) detection. The cognitive ability of AD individuals displays significant variation compared to healthy individuals [21].

Winer and Frankenberg et al. [9, 10] demonstrated the relevance of linguistic features such as parts-of-speech (POS) and word categories to AD. More recently, the development of pre-trained models has mitigated the challenge in data-scarce tasks, leading to extensive research on deep embeddings for AD detection, encompassing speech-based [11–14], text-based [15–18], and multi-modal approaches [11, 19, 20].

Beyond the challenge of feature extraction and representation learning, we propose that AD detection faces unique difficulties inherent to the nature of Alzheimer’s Disease. Medical literature [21–23] establishes AD as a *degenerative disorder characterized by a continuum of pathophysiological changes, resulting in gradual cognitive and functional decline*. This spectrum of cognitive performance among AD patients is illustrated in Fig. 1(a). Consequently, under standard classification modeling, samples with the same label may exhibit varying degrees of cognitive impairment patterns, which could lead to considerable variability in recognition features. This distinguishes AD detection from classification tasks such as image classification, which typically involve more consistent features under clearly defined categories. An ideal solution would be to model AD detection by regression or multi-way classification to capture this continuum. However, granular labels are rarely provided for most cognitive assessment tasks, especially



(a) Sample Score Estimator (b) Soft Target Distillation (SoTD) (c) Instance-level Re-balancing (InRe)

Figure 2: Illustration of the proposed approaches. $KL(\cdot)$ represents the KL divergence function. C_1 to C_n represent component classifiers.

for open-source datasets. While reference continuous scores are occasionally available (such as Mini-Mental State Examination scores), these scores are obtained from different tests that are distinct from the assessment in the data and can hardly be regarded as gold standard for supervision. Therefore, binary classification remains the prevailing method for Alzheimer’s Disease (AD) detection.

This work takes a first step towards the within-class variation (WCV) issue in AD detection. We inspect the sample variability beyond binary labels using the English ADReSS [24] and Cantonese CU-MARVEL [6] datasets (see Fig. 1 (b) and Fig. 1 (c)). Even with comparable class members, AD samples functioning a larger variance in cognitive function compared to healthy samples. Given these findings, we propose that the conventional binary classification paradigm overlooks two critical aspects: a) *within-class heterogeneity (WCH)*: samples with varying AD severity are assigned to the same class, potentially inhibiting the model’s sensitivity to certain changes in cognitive function; and b) *instance-level imbalance (ILI)*: the frequency of varying-severity samples is imbalanced even with balanced class size, thereby introducing potential bias. A key challenge in tackling these issues is the lack of informative instance-level labels indicating sample severity. Accordingly, this work explored sample score estimation that distills proxy soft labels from ensemble models *with only hard label supervision*. Building upon sample score estimation, we subsequently propose two approaches addressing WCH and ILI respectively: Soft Target Distillation (SoTD) and Instance-level Re-balancing (InRe). SoTD approach leverages the label refinery approach [25] to train classifiers using informative soft label supervision. And InRe approach re-weights imbalanced samples at instance level using log posterior ratios and inverse kernel density.

We analyzed the proposed methods using the ADReSS and CU-MARVEL corpora. The experimental results validated that the estimated sample scores aligned with the corresponding cognitive scores, despite the fact that cognitive scores were not available during training. In addition, the InRe method guides model training to focus more on under-represented AD instances. Finally, the proposed strategies exhibit remarkable performance improvements on both evaluation datasets. This work presents an early investigation of the within-class variation issue in AD detection. We seek to provide helpful insights for developing more robust and reliable AD detection models.

2. Approaches

We first revisit AD detection as a binary classification task. Given an input feature z derived from sample x , the posterior probability p_+ and p_- of x being positive (AD) of negative (Non-AD) are estimated by a neural network classifier $c(z; \theta)$, optimized using binary cross-entropy (CE) loss $L_{BCE} = -\sum(H(y_-, p_-) + H(y_+, p_+))$, where y stands for ground-truth AD label.

2.1. Sample score estimation

Sample score estimation is designed to quantify within-class heterogeneity and imbalance for subsequent modules. This module estimates informative sample-specific soft scores that serve to conditionalize binary labels. We hypothesized that classification models are implicitly able to measure and rank samples based on pattern similarities [26]. Consequently, in this module vanilla binary classifiers were utilized to generate sample-wise posterior probabilities, which act as soft scores. To mitigate the randomness during training, we adopt an ensemble approach that averages the predictions of a series of component models. This component is illustrated in Fig. 2 (a).

2.2. Soft target distillation

To tackle the issue of within-class heterogeneity, we aim for the model to be sensitive to the subtle differences within classes. We accomplish this through soft target distillation (SoTD). In this approach, a subsequent classifier is trained using only the soft targets obtained from the sample score estimation. This concept draws inspiration from label refinery [25], which was originally proposed in image classification to deal with cases when one-hot labels cannot cover multiple objectives in a single image. Specifically, the posterior probabilities $p(\hat{y}_1), p(\hat{y}_2)$ generated by the sample score estimator were employed to supervise a new classifier $c'(z; \theta)$ by minimizing the following KL-divergence:

$$L_{KL} = p(y_1) \log\left(\frac{p(y_1)}{p(\hat{y}_1)}\right) + p(y_2) \log\left(\frac{p(y_2)}{p(\hat{y}_2)}\right) \quad (1)$$

where $p(y_1)$ and $p(y_2)$ are the output of $c'(z; \theta)$. It is important to note that $c'(z; \theta)$ did not see the original hard label y . This multi-stage approach has been demonstrated to be effective in addressing long-tailed recognition problems, as in [27]. Moreover, an alternative approach would be to combine a soft target with standard CE loss for distillation. However, our preliminary experiments and prior work [25] suggest that this does not further benefit the performance.

2.3. Instance-level re-balancing

Class re-balancing is a category of methods to address class imbalance. It typically involves emphasizing and de-emphasizing specific classes through re-sampling or loss re-weighting. In this work, we adapt this concept to the instance level to deal with within-class imbalance. The key challenge of instance-level re-balancing is to cluster samples according to AD severity to obtain a frequency distribution. We accomplish this through the following steps: First, for each sample, we measure the scaled AD confidence using the log-probability ratio $l_x = \log(p_+/p_-)$. Second, we calculate $L_X = \{l_{x_0}, \dots, l_{x_n}\}$ for all n samples in training data. Then, we apply density estimation with a Gaussian kernel $K(l_x, l'_x)$ to obtain probability densities representing sample frequencies

$\tilde{p}(x)$:

$$\tilde{p}(x) \triangleq \hat{p}(l_x) = \int_{L_x} K(l_x, l'_x) p(l'_x) dl'_x \quad (2)$$

where $p(l'_x)$ denotes the bin frequencies calculated during density estimation. It is important to note that the bandwidth is an important hyperparameter, as it controls the variance of $p(l'_x)$ across samples, thereby controlling the sharpness of final sample weights distribution. Finally, the inverse sample frequency $\tilde{p}_{inv}(x) = -\ln(\tilde{p}(x))$ was used as weights, which are multiplied by sample losses to re-balancing sample contributions during model training. This entire process is graphically depicted in Fig. 2 (c).

3. Experimental setup

3.1. Dataset

ADReSS This is a frequently utilized dataset for AD detection, which is derived from the ADReSS challenge [24]. It encompasses speech recordings and manual transcriptions obtained from 156 participants engaged in the Cookie Thief picture description task [2]. The dataset is partitioned into a training set consisting of 108 samples and a test set with 48 samples, and there is an equal distribution of positive and negative cases within these subsets.

CU-MARVEL This is a Cantonese corpus that was developed for the study of neurocognitive disorder diseases. It is composed of speech recordings from a sequence of cognitive assessment tasks. In this particular work, we specifically employed the data from the Rabbit Story task [4] within this corpus. Because this task is centered around the spontaneous speech of the participants. We designate this subset as the *CUMV-R* dataset. The CUMV-R dataset contains manual speech transcriptions from 385 participants, among which 153 samples are positive and 232 samples are negative.

3.2. AD detection model

Pre-trained language models have demonstrated remarkable performance in AD detection [6, 28, 29]. In this study, BERT-family models [30, 31] were employed to extract linguistic features from transcribed speech data. Subsequently, multi-layer perception (MLP) classifiers and the Cross-entropy loss function were utilized. For ADReSS corpus, we emulated the work [28] used bert-base-uncased model¹ as the feature extractor. The classifier contains 2 hidden layers with sizes of (32, 16). It was optimized using Adam optimizer with a learning rate of $1e-3$, a batch size of 16, and the training was carried out for 20 epochs. Regarding this, we followed the work in [6] and used a Chinese RoBERTa model² [32] as the feature extractor. The classifier had similar configurations, but the hidden layer sizes were set to (64, 16) and the learning rate was $1e-4$.

Along with the vanilla AD detection models, we implemented two baseline systems. Unlike SoTD, we utilized model ensembling, where the posteriors of the model outputs were averaged for decision fusion. Compared to the InRe method, we employed resampling-based class re-balancing, with the class sample rate set as the reciprocal of class frequencies.

3.3. Proposed approaches

We incorporated 5 component models in the sample score estimator and the model ensembling baseline. As for density estimation in the InRe approach, we set the bin sizes as 2 by default.

¹<https://huggingface.co/google-bert/bert-base-uncased>

²<https://huggingface.co/hfl/chinese-roberta-wwm-ext>

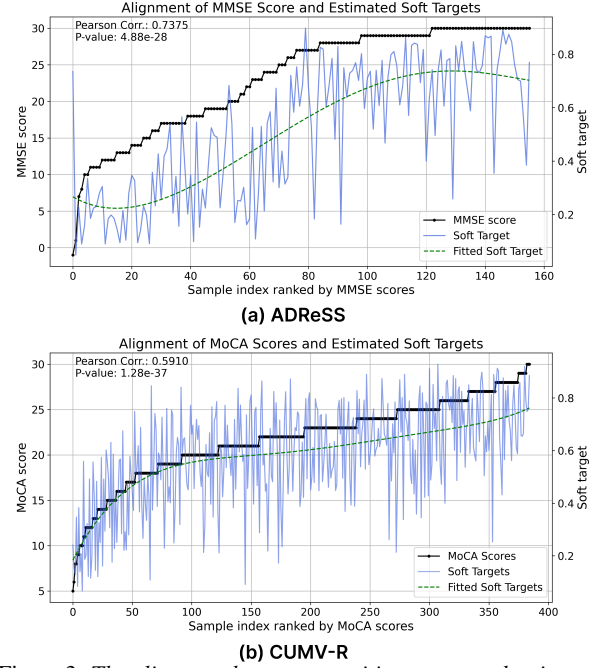


Figure 3: The alignment between cognitive scores and estimated soft targets in the ADReSS (a) and CUMV-R (b) datasets.

4. Results and discussions

4.1. Validate estimated sample scores

An effective sample score estimation serves as a prerequisite for modeling within-class heterogeneity and imbalance using SoTD and InRe approaches. We first generate estimated scores for every sample in two datasets and observe their correlation with participants’ cognitive scores (i.e., MMSE scores in ADReSS; MoCA scores in CU-MARVEL). It is important to note that the cognitive scores referenced here should be viewed as assistant labels that indicate participants’ cognitive ability in certain aspects, whereas the hard labels represent ground-truth labels for AD. As drawn in Fig. 3, we can observe that although the cognitive score labels were not seen during training, the estimated sample scores exhibit alignment with cognitive scores. Their Pearson Correlation is 0.7375 and 0.5910 in the two datasets respectively. Correspondingly, as used in the proposed SoTD method, guiding classifier training with these soft targets would take into account more comprehensive cognitive abilities of participants compared to relying on binary hard labels. Although we found some soft targets show “noisy” deviations from cognitive scores, we contend that the model overall benefits from these soft targets rather than just binary labels with values of 1 and 0.

4.2. AD detection performance

Table 1 presents our experimental results on ADReSS and CUMV-R corpus. To reduce randomness, we performed 10-fold cross-validation for 20 random runs and reported the averaged results. On the ADReSS corpus, the SoTD approach outperformed the baseline and ensemble systems across all 4 metrics. This demonstrates general improvement in AD classification. The InRe approach also led to improvement in the balanced accuracy and F1 metrics, while remarkably enhancing the recall rates. This aligns with our expectation because the InRe approach could emphasize the infrequent positive (AD) samples in the dataset (as shown in Fig. 1 (b)), thereby improving the model’s sensitivity. It is worth noting that *this contribution*

Table 1: 10-fold cross-validation results for AD detection using the ADReSS dataset. Each result is an average of 20 random runs. “B-Acc.” refers to balanced accuracy, and asterisk (*) indicates the proposed methods.

	ADReSS			
	B-Acc.	F1	Prec.	Recall
Baseline	.8330 \pm .096	.8202 \pm .115	.8371	.8273
Ensemble	.8399 \pm .105	.8246 \pm .113	.8474	.8271
SoTD*	.8608 \pm .104	.8396 \pm .106	.8602	.8382
InRe*	.8505 \pm .107	.8351 \pm .113	.8286	.8573

	CUMV-R			
	B-Acc.	F1	Prec.	Recall
Baseline	.6278 \pm .067	.5072 \pm .110	.5993	.4615
Ensemble	.6266 \pm .070	.5051 \pm .113	.5983	.4568
SoTD*	.6356 \pm .065	.5323 \pm .090	.5893	.4978
Resampling	.6316 \pm .074	.5330 \pm .142	.5546	.5640
InRe*	.6407 \pm .074	.5623 \pm .099	.5518	.5881

differs from conventional class re-balancing methods since the number of positive and negative samples is equal in the training set. Thus, we contend that the advantage of the InRe approach stems its awareness of within-class imbalance. The resampling approach was not carried out in this dataset, as this dataset is already balanced and the results will be the same as the baseline.

The CUMV-R dataset poses a more challenging scenario for several reasons: (a) it contains more variations in the cognitive scores; (b) positive and negative samples are more imbalanced (ratio=1.51); and (c) Cantonese BERT models might not be as well-developed as English models. This is particularly evident in the low recall rates. The SoTD approach outperform both baseline and ensembling methods by balancing the model’s predictions, evidenced by the improved recall rate. The resampling and InRe methods both show further improvement in recall rate, while the InRe method improved the models’ sensitivity by a large margin (0.4615 \rightarrow 0.5881) and therefore remarkably improved the overall F1 score (0.5072 \rightarrow 0.5623).

4.3. Inspect sample weights in InRe approach

To gain deeper insight into the InRe approach, a natural question is how sample weights are assigned across samples with varying cognitive levels. In Fig. 4, we collected all calculated sample weights and grouped them by corresponding cognitive score. Again, cognitive scores are not accessible during the training phase. On the ADReSS corpus, the weights roughly correlate with the MMSE scores, but not that well. We attribute this to the dispersion of sample groups in this dataset, where most groups contain fewer than 10 samples and can hardly guarantee statistical stability. Nevertheless, samples with MMSE scores below 15 tend to be assigned notably higher weights. In contrast, samples in the CUMV-R dataset display more distinct negative correlations – the less frequent groups are assigned larger weights. These findings imply that the InRe approach attempts to regulate the models’ training to focus more on AD instances within sparse groups. This behavior could explain the improvement in model sensitivity as presented in Table 1.

4.4. Effect of bandwidths in InRe

Bandwidth is an important hyperparameter within the InRe approach during density estimation. It controls estimation variance and bias, thereby affecting the sharpness of the sample weights distribution. Table 2 presents a comparison of detection

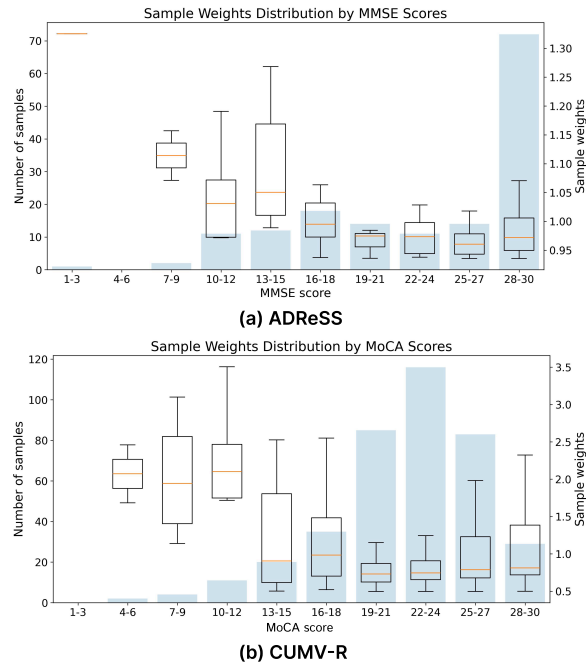


Figure 4: Distribution of assigned sample weights across cognitive scores in the InRe method.

performance attained by employing different bandwidths. We found that using relatively *smaller bandwidths* generally results in better performance, while the optimal result is obtained using a bandwidth of 2. This could be because smaller bandwidths preserve more disparities among samples, while too small bandwidths (i.e., Bandwidth=1) incorporate potential noise in sample score estimation.

Table 2: A comparison of different bandwidths adopted in the InRe approach on the CUMV-R dataset. “B-Acc.” refers to balanced accuracy.

	Bandwidths					
	1	2	4	8	16	64
B-Acc.	.6357	.6407	.6330	.6367	.6303	.6241
F1	.5582	.5623	.5528	.5607	.5524	.5255
Precision	.5459	.5518	.5412	.5450	.5398	.5508
Recall	.5867	.5881	.5804	.5925	.5800	.5169

5. Conclusions

This work explores the issue of within-class variation in Alzheimer’s Disease (AD) detection. We posit that binary classification may overlook two crucial aspects: within-class heterogeneity (WCH) and instance-level imbalance (ILI). We further introduce two simple yet effective methods to address these problems: soft target distillation (SoTD) and instance-level re-balancing (InRe). Experiments on ADReSS and CUMARVEL corpora demonstrated their advantages in detection performance. Future work will explore the combination of the SoTD and InRe methods, and investigate the incorporation of cognitive scores during model training.

6. Acknowledgements

This work is supported by the HKSARG Research Grants Council’s Theme-based Research Grant Scheme (Project No. T45-407/19N) and the CUHK Stanley Ho Big Data Decision Research Centre.

7. References

- [1] C. Lynch, "World alzheimer report 2019: Attitudes to dementia, a global survey: Public health: Engaging people in adrd research," *Alzheimer's & Dementia*, 2020.
- [2] E. Giles, K. Patterson, and J. R. Hodges, "Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information," *Aphasiology*, vol. 10, no. 4, pp. 395–408, 1996.
- [3] H. Goodglass, E. Kaplan, and B. Barresi, *BDAE-3: Boston Diagnostic Aphasia Examination-Third Edition*. Lippincott Williams & Wilkins Philadelphia, 2001.
- [4] J. Reilly, M. Losh, U. Bellugi, and B. Wulfeck, "'frog, where are you?' narratives in children with specific language impairment, early focal brain injury, and williams syndrome," *Brain and language*, vol. 88, no. 2, pp. 229–247, 2004.
- [5] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article," *Frontiers in Psychology*, vol. 12, p. 620251, 2021.
- [6] H. Meng, B. Mak, M.-W. Mak, H. Fung, X. Gong, T. Kwok, X. Liu, V. Mok, P. Wong, J. Woo *et al.*, "Integrated and enhanced pipeline system to support spoken language analytics for screening neurocognitive disorders," *Interspeech*, 2023.
- [7] J. Kang, J. Li, J. Li, X. Wu, and H. Meng, "Not all errors are equal: Investigation of speech recognition errors in alzheimer's disease detection," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 254–258.
- [8] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," 2017.
- [9] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, 2016.
- [10] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *ASRU*. IEEE, 2019.
- [11] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," in *INTERSPEECH*, 2020.
- [12] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 624137, 2021.
- [13] J. Li, K. Song, J. Li, B. Zheng, D. Li, X. Wu, X. Liu, and H. Meng, "Leveraging pretrained representations with task-related keywords for alzheimer's disease detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [14] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection," in *Interspeech*, vol. 2021. NIH Public Access, 2021, p. 3790.
- [15] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.
- [16] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," in *INTERSPEECH*, 2020.
- [17] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, 2021.
- [18] Y. Wang, T. Wang, Z. Ye, L. Meng, S. Hu, X. Wu, X. Liu, and H. Meng, "Exploring linguistic feature and model combination for speech recognition based automatic ad detection," *arXiv preprint arXiv:2206.13758*, 2022.
- [19] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling," *IEEE Access*, 2021.
- [20] J. Li, Y. Wang, J. Li, J. Kang, B. Zheng, S. Wong, B. Mak, H. Fung, J. Woo, M.-W. Mak *et al.*, "Detecting neurocognitive disorders through analyses of topic evolution and cross-modal consistency in visual-stimulated narratives," *arXiv preprint arXiv:2501.03727*, 2025.
- [21] D. K. Johnson, M. Storandt, J. C. Morris, and J. E. Galvin, "Longitudinal study of the transition from healthy aging to alzheimer disease," *Archives of neurology*, vol. 66, no. 10, pp. 1254–1259, 2009.
- [22] P. S. Aisen, J. Cummings, C. R. Jack, J. C. Morris, R. Sperling, L. Frölich, R. W. Jones, S. A. Dowsett, B. R. Matthews, J. Raskin *et al.*, "On the path to 2025: understanding the alzheimer's disease continuum," *Alzheimer's research & therapy*, vol. 9, pp. 1–10, 2017.
- [23] B. Dubois, H. H. Feldman, C. Jacova, J. L. Cummings, S. T. DeKosky, P. Barberger-Gateau, A. Delacourte, G. Frisoni, N. C. Fox, D. Galasko *et al.*, "Revising the definition of alzheimer's disease: a new lexicon," *The Lancet Neurology*, vol. 9, no. 11, pp. 1118–1127, 2010.
- [24] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," *INTERSPEECH*, 2020.
- [25] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving imagenet classification through label progression," *arXiv preprint arXiv:1805.02641*, 2018.
- [26] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3903–3911.
- [27] T. Li, L. Wang, and G. Wu, "Self supervision to distillation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 630–639.
- [28] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for alzheimer's disease detection," in *ICASSP*. IEEE, 2021, pp. 6423–6427.
- [29] Y. Wang, T. Wang, Z. Ye, L. Meng, S. Hu, X. Wu, X. Liu, and H. Meng, "Exploring linguistic feature and model combination for speech recognition based automatic ad detection," *arXiv preprint arXiv:2206.13758*, 2022.
- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019.
- [31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.
- [32] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. Online: Association for Computational Linguistics, Nov. 2020, pp. 657–668. [Online]. Available: <https://www.aclweb.org/anthology/2020.findings-emnlp.58>