SPZ: A Semantic Perturbation-based Data Augmentation Method with Zonal-Mixing for Alzheimer's Disease Detection

Fangfang Li[1,2], Cheng Huang[1], Puzhen Su[1]*, Jie Yin[3]
[1]School of Computer Science and Engineering, Central South University, China
[2]Xiangjiang Laboratory, Changsha, Hunan province, China
[3]University of Sydney, Sydney, Australia
{lifangfang,224712158,214711091}@csu.edu.cn, jie.yin@sydney.edu.au

Abstract

Alzheimer's Disease (AD), characterized by significant cognitive and functional impairment, necessitates the development of early detection techniques. Traditional diagnostic practices, such as cognitive assessments and biomarker analysis, are often invasive and costly. Deep learning-based approaches for non-invasive AD detection have been explored in recent studies, but the lack of accessible data hinders further improvements in detection performance. To address these challenges, we propose a novel semantic perturbation-based data augmentation method that essentially differs from existing techniques, which primarily rely on explicit data engineering. Our approach generates controlled semantic perturbations to enhance textual representations, aiding the model in identifying AD-specific linguistic patterns, particularly in scenarios with limited data availability. It learns contextual information and dynamically adjusts the perturbation degree for different linguistic features. This enhances the model's sensitivity to AD-specific linguistic features and its robustness against natural language noise. Experimental results on the ADReSS challenge dataset demonstrate that our approach outperforms other strong and competitive deep learning methods.

1 Introduction

Alzheimer's Disease (AD) casts a shadow over the globe as an affliction of the mind, eroding cognitive and functional capabilities with devastating thoroughness (Roark et al., 2011; Deture and Dickson, 2019). As the condition progresses, the irrevocable nature of this disorder becomes markedly pronounced, underscoring the imperative for early detection and opportune intervention. Contemporary medical diagnostics have predominantly hinged on cognitive testing and bio-marker analysis (Guo et al., 2020), yet these approaches are notably time-intensive and financially burdensome, thereby hindering their widespread adoption for potential patients. Recent advancements in automated speech and text analysis offer a more cost-effective and scalable alternative for AD detection (Prabhakaran et al., 2018). Patients with AD exhibit distinctive linguistic patterns (Liu et al., 2022), characterized by an increased reliance on filler words, a paucity of informative nouns and verbs, disorganized syntax, alongside heightened frequencies of pauses and hesitations(Yuan et al., 2020a; Greta et al., 2015). These linguistic features provide promising potential for language-based analysis mechanisms in detecting AD, indicating that enhanced sensitivity in capturing these patient-specific features could improve the efficiency of early-stage detection and intervention.

With the advancements of deep learning, many neural network-based methods (Ilias and Askounis, 2023; Koo et al., 2020; Zhu et al., 2021) have been proposed to detect AD. Yet, the primary challenges obstructing the development of effective detection lie in the limited sensitivity of existing methods to the unique linguistic features of AD patients and the shortage of labeled data. The difficulty in data collection (Chen et al., 2023), the high costs associated with data labeling, and privacy concerns have all hindered the availability of accessible and labeled linguistic datasets. This scarcity increases the risk of model overfitting and impedes the exploration of specific features that are highly relevant to AD patients.

In recent years, Data Augmentation (DA) has been proposed as an effective technique to alleviate the scarcity of available datasets and improve model generalization. Previous DA methods for AD detection (Guo et al., 2021; Roshanzamir et al., 2021; Liu et al., 2021) primarily involve performing explicit transformations of the original text sequences (e.g., random deletion or lexical substi-

*Corresponding author

tution etc.) (Novikova, 2021; Duan et al., 2023; Hlédiková et al., 2022), paraphrasing (Cai et al., 2023), or text generation (Guo et al., 2020). Although these methods alleviate the problem of data availability to a certain extent, they might result in augmentation bias and impair the semantics of the original texts. First, the randomly deleted AD samples may contain AD-specific features removed, which are indeed semantically closer to non-AD samples. Second, random operations have difficulty in preserving AD-specific features that are crucial for AD diagnosis. As a consequence, the model might be misled to learn inappropriate pseudo AD patterns, thereby hurting the model's generalization ability. In addition, various types of DA methods require varying degrees of explicit data engineering, thus limiting the portability and applicability of these methods.

To address the above issues, we aim to develop an effective Data Augmentation (DA) method to improve the performance of AD detection in data-limited scenarios. Our approach centers on utilizing semantic perturbations to enhance the sensitivity of AD detection models in capturing AD-specific features while retaining the robustness to natural noise that may appear in user-generated texts. In our work, we propose a *Semantic Perturbation with Zonal-mixing* (SPZ) DA framework. SPZ consists of a Generator module and a Mediator module. The Generator automatically augments the given text with varying degrees of noise generated from a Gaussian distribution, optimized using a temperature-dependent Gumbel-Softmax trick (Jang et al., 2017). Unlike Generative Adversarial Networks (GANs) that typically create new data samples, SPZ focus on adding controlled perturbations to the existing data samples. To control the degree to which different types of words are perturbed, the Mediator assesses all initial word representations, assigning corresponding perturbation probabilities. Finally, all initial word representations and generative noise representations are mixed to produce a refined representation as input to the final AD classifier. This dual-module approach ensures the preservation of the text's original semantics with subtle modifications to improve the effectiveness of model training. We illustrate that our proposed approach achieves competitive performance on the ADReSS challenge dataset compared to strong baselines.

Our contributions are summarized as follows:

1. To the best of our knowledge, we are the first to propose a perturbation-based semantic augmentation approach for the detection of AD transcripts without manual data engineering.

2. We propose SPZ, a novel plug-and-play DA framework with semantic perturbations, which perturbs different features with varying degrees of noise, to enhance the efficacy of AD detection in data-limited scenarios.

3. Experimental results on the ADReSS challenge dataset demonstrate the effectiveness and competitiveness of our method.

2 The ADReSS Challenge Dataset

We utilize the ADReSS dataset (Luz et al., 2020), which serves as a standardized benchmark for the Alzheimer's Disease (AD) research community, facilitating research in AD detection. This dataset comprises 156 speech samples with corresponding transcripts from participants—both non-AD (35 male, 43 female) and AD-affected (35 male, 43 female) English speakers—performing the Cookie Theft picture description task (Giles et al., 1996). Our study focuses exclusively on the transcript data, formatted according to the CHAT protocol (MacWhinney, 2000), a recognized standard for TalkBank data. As shown in Table 1, the dataset has a balanced distribution across age and gender, aiming to reduce potential bias in prediction tasks.

Age	**AD (train / test)**		**Non-AD (train / test)**	
	Male	**Female**	**Male**	**Female**
[50, 55)	1 / 1	0 / 0	1 / 1	0 / 0
[55, 60)	5 / 2	4 / 2	5 / 2	4 / 2
[60, 65)	3 / 1	6 / 3	3 / 1	6 / 3
[65, 70)	6 / 3	10 / 4	6 / 3	10 / 4
[70, 75)	6 / 3	8 / 3	6 / 3	8 / 3
[75, 80)	3 / 1	2 / 1	3 / 1	2 / 1
Full set	24 / 11	30 / 13	24 / 11	30 / 13

Table 1: Statistics of the ADReSS dataset with age and gender details

The ADReSS dataset's limited size, resulting from data collection difficulties and privacy concerns, poses a significant challenge for developing robust AD classification models. To tackle this challenge, we propose a novel data augmentation method aimed at enhancing textual representations in data-limited scenarios. Our strategy involves generating controlled semantic perturbations, de-

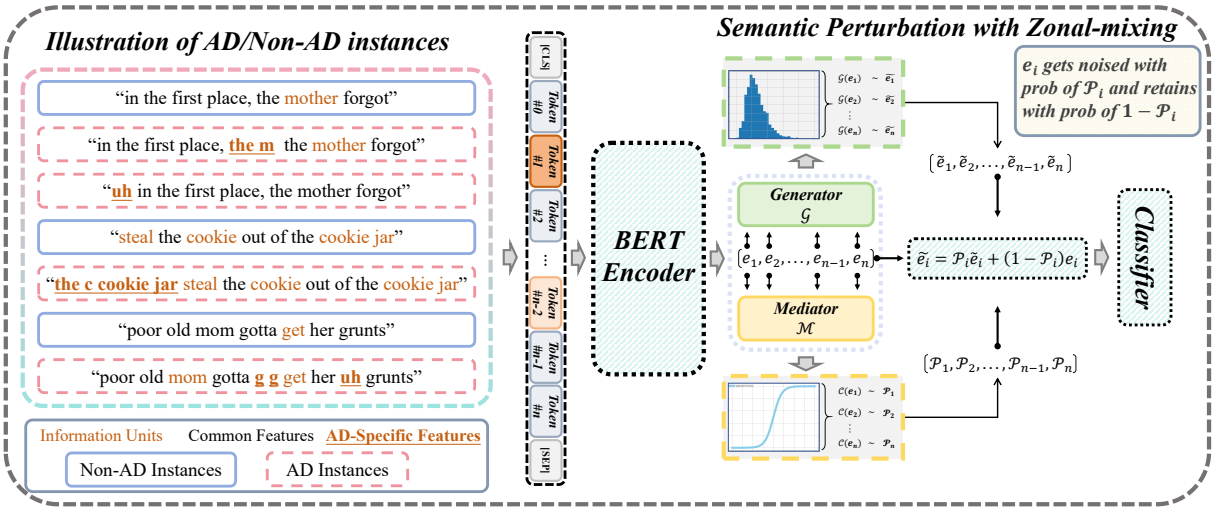


Figure 1: Illustration of our proposed Semantic Perturbation with Zonal-mixing (SPZ) framework

signed to boost model performance and reliability in AD detection.

3 Methods

Our proposed method aims to generate better representations of Alzheimer's Disease (AD) transcripts in data-limited scenarios, which can be easily optimized and utilized to perform effective early-stage AD detection. In our framework, each transcript is initially embedded into a pre-trained language model BERT (Devlin et al., 2019) at the token level, thereby obtaining representations for each word. These word representations are subsequently channeled through three distinct information pathways: 1) Fed into a Generator module, which is designed to derive semantic perturbations; 2) Transmitted into a Mediator module, which controls the corresponding degrees of perturbation for each word; 3) Combined with the generative word semantic perturbations and then input into the final AD classifier. Details of our proposed approach are presented in subsequent subsections.

3.1 Text Encoder

Building upon prior studies (Duan et al., 2023), our method adopts the pre-trained language model BERT[1] (bert-base-uncased) as our text encoder. Unlike current BERT-based AD detection methods that directly use the [CLS] token for classification, we use a token-level embedding strategy to capture more detailed and fine-grained feature information.

As shown in Figure 1, any input transcript $T = \{t_1, t_2, \cdots, t_n\}$, comprising various features (i.e., information units, common features and AD-specific features), is processed through the BERT encoder:

$$\mathbf{E} = \text{BERT}(T) \tag{1}$$

By leveraging the BERT's Transformer architecture, we can capture interactions among different types of features and obtain the contextualized sentence representation $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}$, which consists of n token-level representations. To best utilize the latent representations, we use all token-level representations, excluding the special tokens [CLS] and [SEP], for subsequent perturbation-based DA module.

3.2 Semantic Perturbation with Zonal-mixing

To enhance effectiveness and enable plug-and-play use in data-limited scenarios, we propose **S**emantic **P**erturbation with **Z**onal-mixing (**SPZ**) , a new method focused on augmenting textual representations via semantic perturbation. SPZ aims to refine the model's sensitivity to AD-specific linguistic features and improve its robustness against natural linguistic variations.

Perturbation Generator This module is designed to generate perturbations with more variability, encouraging the model to uncover and adapt to diverse linguistic patterns, while avoiding the semantic bias derived from random and direct textual changes. For each given transcript $T = \{t_1, t_2, \cdots, t_n\}$, we utilize the BERT encoder to obtain the contextualized representations $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}$ at the token level, where $\mathbf{e}_k$, $1 \leq k \leq n$, indicates an h-dimensional representation vector of token k.

[1] https://github.com/google-research/bert

For all token representations, we generate the corresponding n noise representations from a Gaussian distribution:

$$\tilde{\mathbf{U}} = \{\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \cdots, \tilde{\mathbf{u}}_n\} \sim \mathcal{N}(0, \sigma^2), \quad (2)$$

where $\tilde{\mathbf{u}}_k \in \mathcal{R}^{1\times h}$, $1 \leq k \leq n$, indicates the noise perturbed to token representation $\mathbf{e}_k$, and σ is the standard variance[2]. As perturbations sampled from a Gaussian distribution typically lack an explicit direction in the semantic space, the augmentation performance of injecting uniformly sampled perturbations into token-level representations may not be insufficient.

To introduce more significant perturbations into specific elements of the representations while keeping others relatively unchanged, we transform the initial noise representations into a skewed distribution to derive the Gumbel noise $\mathbf{g}$, given by:

$$\mathbf{g} = -\log(-\log(\tilde{\mathbf{u}} + \varepsilon) + \varepsilon). \quad (3)$$

Based on the above, we adopt the Gumbel-Softmax trick (Jang et al., 2017) with low temperature to obtain the perturbed representations $\mathbf{E}' = \{\mathbf{e}'_1, \mathbf{e}'_2, \cdots, \mathbf{e}'_n\}$ via:

$$\mathbf{e}'_i = \frac{\exp\{(\mathbf{e}_i + \mathbf{g}_i)/\tau\}}{\sum_{j=1}^{n} \exp\{(\mathbf{e}_j + \mathbf{g}_j)/\tau\}}, \quad (4)$$

where $\mathbf{e}_k \in \mathcal{R}^{1\times h}$ and τ represents the temperature parameter that controls the smoothness of the perturbed distribution, and we set $\tau = 0.2$ in our experiment. With a lower τ, the perturbations injected into specific words become more extreme and targeted, resulting in a less smooth distribution. This is critical for accurately detecting key linguistic patterns associated with AD.

Perturbation Mediator To achieve auto-balancing during training, we design the perturbation Mediator to model the perturbation distribution from the initial contextualized representations $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}$ as:

$$p_i = \frac{1}{1 + \exp\{-(\mathbf{w}_i \cdot \mathbf{e}_i + b)\}}, \quad (5)$$

where $\mathbf{w}_i \in \mathcal{R}^{1\times h}$ and $b_i \in \mathcal{R}^{1\times 1}$ are trainable weights matrix and bias vector. $p_i \in (0, 1)$ indicates the degree to which $\mathbf{e}_i$ is perturbed. The sigmoid function used in Eq. 5 ensures a smooth and continuous transition between the original and perturbed states, providing fine-grained control over the perturbation degree.

Following the computation of perturbation degrees, we employ a zonal-mixing strategy that integrates the original and perturbed representations to obtain the final mixed-up representations $\tilde{\mathbf{E}} = \{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \cdots, \tilde{\mathbf{e}}_n\}$, given by:

$$\tilde{\mathbf{e}}_i = (1 - p_i)\mathbf{e}_i + p_i\mathbf{e}'_i. \quad (6)$$

The combination of the original and perturbed representations enables the model to learn flexibly from both clean and perturbed samples. Furthermore, since the perturbation degree in the Mediator module is transparent and interpretable, it offers a bridge between the black-box and white-box models, providing a potential perspective for credible AD detection. We will further discuss the credible detection results in Section 4.4.

3.3 Alzheimer's Disease Classifier

After deriving the mixed-up representations $\tilde{\mathbf{E}} = \{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \cdots, \tilde{\mathbf{e}}_n\}$, we apply a global max pooling layer to reduce the dimensionality, followed by a fully connected layer for the final classification:

$$\begin{aligned} \mathbf{z} &= \text{GlobalMaxPooling}(\tilde{\mathbf{E}}), \\ \hat{y} &= \text{Sigmoid}(\mathbf{z}), \end{aligned} \quad (7)$$

where $\mathbf{z} \in \mathcal{R}^{1\times h}$ and $\hat{y} \in (0, 1)$ indicates the probability of predicting AD. The Binary Cross Entropy loss is used as the loss function $\mathcal{L}_{BCE}$ for model training:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \quad (8)$$

where y_i denotes the ground-truth label of the input sentence, respectively.

4 Experiments

We implement our proposed method using bert4keras (Su, 2020) on NVIDIA Tesla V100s with 32GB RAM. Accuracy is used as the primary performance metric due to the well-balanced nature of the ADReSS challenge dataset. Additionally, precision, recall and F1 scores with respect to the AD class are considered to offer a comprehensive evaluation. We use 10-fold cross-validation on the training dataset to assess generalization error, and report the average evaluation metrics across 5 different seeds. The learning rate is set to $2e-05$ and batch size is set to 8.

[2]In our work, we set $\sigma = 1$ and $h = 786$ denotes the hidden size of BERT.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
LDA (Luz et al., 2020)	75.0	83.0	62.0	71.0
CNN (Meghanani et al., 2021)	72.9	72.0	75.0	73.0
fastText (Meghanani et al., 2021)	83.3	86.0	79.0	83.0
BERT (Balagopalan et al., 2021)	83.3	83.9	83.3	83.3
Fusion (Campbell et al., 2021)	83.3	80.1	87.5	84.0
BERT (BT DE) (Hlédiková et al., 2022)	84.0	-	81.1	-
SVM (BT RU) (Hlédiková et al., 2022)	85.0	-	79.0	-
CDA$_{\text{single}}$ (Duan et al., 2023)	87.5	88.1	83.3	86.9
SPZ$_{\text{single}}$ (ours)	**90.0**	**88.8**	**91.7**	**90.2**
Ensemble (Sarawgi et al., 2020)	83.0	83.0	83.0	83.0
ERNIE0p (Yuan et al., 2020b)	85.4	94.7	75.0	83.7
ERNIE3p(Yuan et al., 2020b)	89.6	95.2	83.3	88.9
CDA$_{\text{ensemble}}$ (Duan et al., 2023)	91.7	**100.0**	83.3	90.9
SPZ$_{\text{ensemble}}$ (ours)	**93.8**	92.0	**95.8**	**93.9**

Table 2: Predictive performance of each comparing method on the ADReSS challenge dataset. Accuracy, Precision, Recall and F1 score (F1) are used as evaluation metrics.

4.1 Baseline Methods

We compare our method with two sets of competitive AD detection methods including: single model methods and ensemble methods.

The single model methods include (1) LDA(Luz et al., 2020), as the baseline of the ADReSS challenge, which utilizes linear discriminant analysis to detect AD; (2) Feature-based methods (Meghanani et al., 2021) that employ fastText (bi-grams and tri-grams) and a CNN model (bi-grams, tri-grams and 4-grams) to extract n-gram-based linguistic features; (3) BERT (Balagopalan et al., 2021) that takes the [CLS] token as the global representation for classification; (4) Fusion (Campbell et al., 2021) that integrates acoustic and linguistic features to augment the classification performance; (5) SVM (BT RU) and BERT (BT DE) (Hlédiková et al., 2022) employ Back-translation from Russian (RU) and German (DE), respectively, as the textual data augmentation for AD detection.

The ensemble methods include (6) Ensemble (Sarawgi et al., 2020) that implements the majority vote from three individual models; (7) ERNIE0p and ERNIE3p (Yuan et al., 2020b), based on ERNIE-large, which leverage both the original and manually pause-enhanced transcripts for AD detection; (8) CDA (Duan et al., 2023) that proposes a contrastive data augmentation strategy using random deletion as the negative pair and dropout-generated instances as the positive pair.

4.2 Main Results

Our study involves analytical experiments, as shown in Table 2, where we compare our proposed SPZ method against various AD detection techniques, including both single model and ensemble methods.

Single model methods exhibit a diverse range of performance metrics. The baseline LDA yields an accuracy of 75.0%, while the CNN model reports a slightly lower accuracy of 72.9%, suggesting the need for more advanced textual analysis methods for AD detection. Notably, the fastText model yields a substantial increase in recall to 79.0%, underscoring the significance of n-gram features in capturing AD linguistic characteristics. Furthermore, the BERT model, owing to its representation power, achieves a notable accuracy of 83.3%.

Data Augmentation (DA) techniques further enhance the performance of these deep learning models. For instance, Back Translation-based approaches applied to BERT (BT DE) and Support Vector Machines (SVM) (BT RU) demonstrate substantial improvements, achieving an accuracy of 84.0% and 85.0%, respectively. Among the existing methods, CDA$_{\text{single}}$ stands out with a significant accuracy of 87.5% and a precision of 88.1%, showcasing its effectiveness in AD detection.

Our SPZ$_{\text{single}}$ exhibits the best performance over the existing models, achieving an accuracy of 90.0%, along with a precision of 88.8%, recall of 91.7%, and an F1 score of 90.2%. The sta-

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
BERT$_{\text{CLS}}$	83.3	83.3	85.0	83.5
BERT$_{\text{Token}}$	85.0	86.3	84.2	84.9
+ Generator	87.5	88.1	86.7	87.4
SPZ (ours)	**90.0**	**88.8**	**91.7**	**90.2**

Table 3: Ablation study over BERT$_{\text{CLS}}$, BERT$_{\text{Token}}$ and different model components on the ADReSS challenge dataset.

ble precision and significant improvement in recall demonstrate the SPZ's enhanced sensitivity to AD-specific linguistic patterns and its robustness to natural language noise.

Under the ensemble settings, our SPZ$_{\text{ensemble}}$ (majority votes on 5 models) achieves promising performance with an accuracy of 93.8% and an F1 score of 93.9%, surpassing the previous leading CDA$_{\text{ensemble}}$, which has an accuracy of 91.7% and an F1 score of 90.9%. Furthermore, SPZ$_{\text{ensemble}}$ reports a precision of 92.0% and a competitive recall of 95.8%, demonstrating its balanced effectiveness in precisely detecting actual AD cases while maintaining a low rate of misdiagnosed classification.

Type	Example
filler word	*mhm, um, uh, hm, well, yeah, okay etc.*
location	*garage, kitchen, outside, garden, yard etc.*
action	*find, fall, wash, steal, pick, get etc.*
subject	*boys, mom, lady, woman, sister etc.*
object	*dishes, counter, window, stool, curtains, cups, sink, water, cookies, cookie jar etc.*
pronoun	*I, she, he, they, it, that, her, there etc.*
other	*words not in above types.*

Table 4: Seven types of words in the transcripts

4.3 Ablation Study

To verify the contribution of each component in our proposed SPZ framework, we conduct a series of ablation experiments. Table 3 reports the ablation results by comparing our full SPZ model against variants with a specific component ablated.

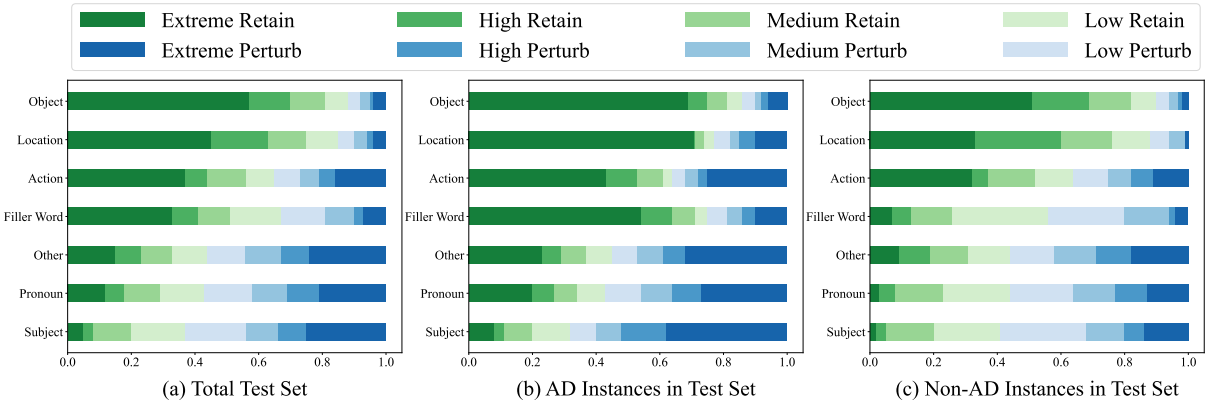
The baseline BERT model, which uses the [CLS] token for classification (BERT$_{\text{CLS}}$), achieves an accuracy of 83.3%, a precision of 83.3%, a recall of 85.0%, and an F1 score of 83.5%. When shifting from the [CLS] token to token-level embeddings (BERT$_{\text{Token}}$), we observe an increment in most metrics, indicating the significance of fine-grained analysis. Adding the Generator module to BERT$_{\text{Token}}$ further improves the performance across all metrics, with particularly notable gains in precision and F1 score, validating the efficacy of semantic perturbation in enhancing model sensitivity to contextual features. Our SPZ, by integrating Generator and Mediator, achieves a remarkable increase in performance with an accuracy of 90.0%, a precision of 88.8%, a recall of 91.7%, and an F1 score of 90.2%. These results markedly outperform the competitive baseline models, highlighting the synergistic effect of the combined modules. Particularly, the significant leap in recall suggests that SPZ is highly effective in identifying AD-related features within the data.

The ablation results clearly demonstrate that each module within the SPZ framework plays a pivotal role in the overall performance. Their seamless integration enhances the model's ability to discern intricate patterns associated with AD, leading to a robust and sensitive AD detection model.

4.4 Analysis of Mediator Module

Deep learning models have long been described as "black-box models" due to the opacity of their prediction processes. The lack of transparency can further lead to distrust in improving the performance of downstream tasks. To illustrate the interpretability of our SPZ method in capturing features crucial for AD detection, we conduct comprehensive case studies of the Mediator module under two distinct scenarios: (1) **AD-specific Features Probing (AFP)**; (2) **Comparison between AD and Non-AD (CAN)**.

To assess the Mediator module's capacity for learning distinct perturbation boundaries, we examine its output for each test instance $\mathbf{S}_i = \{t_1, t_2, \cdots, t_n\}$. This involves statistical analysis of the Perturbation Degree (PD) for each word within the sentences. We categorize words into seven groups based on the transcripts and segment them into eight intervals, each reflecting varying



(a) Total Test Set (b) AD Instances in Test Set (c) Non-AD Instances in Test Set

Figure 2: The distribution of perturbation degree in test dataset (Total, AD instances and Non-AD instances) among seven types of word.

degrees of perturbation as determined by the probabilities generated by the Mediator module. Detailed definitions are provided in Table 4 and Table 5.

Degree	Retention	Perturbation
Low	$(\delta_i - 10\sigma_i, \delta_i]$	$(\delta_i, \delta_i + 10\sigma_i]$
Medium	$(\delta_i - 20\sigma_i, \delta_i - 10\sigma_i]$	$(\delta_i + 10\sigma_i, \delta_i + 20\sigma_i]$
High	$(\delta_i - 30\sigma_i, \delta_i - 20\sigma_i]$	$(\delta_i + 20\sigma_i, \delta_i + 30\sigma_i]$
Extreme	$(0, \delta_i - 30\sigma_i]$	$(\delta_i + 30\sigma_i, 1]$

Table 5: The definition of perturbation degree, where δ_i denotes the average perturbation degree within each sentence S_i, while σ_i represents the perturbation variance within S_i.

Scenario 1: AD-specific Features Probing (AFP) First, we probe the Mediator's sensitivity to AD-specific features. A sensitive Mediator is anticipated to assign lower perturbation degrees to AD-specific features, focusing more on detecting AD. We first analyze the perturbation distributions (see table 5) of seven different word types (see table 4) in the test set. Figure 2(a) provides a detailed view of how the Mediator module perturbs different types of words. For example, clinically important information, such as information units (i.e., task-specific object, location and action) and filter words, show a high retention rate in the overall test set, with high and extreme perturbations being relatively rare. Conversely, features lacking clear distinction (i.e., other, personal pronouns and subjects) exhibit high or extreme perturbation degrees, which indicates the credibility and effectiveness of the Mediator module in identifying key language patterns correlated with AD.

Furthermore, in order to illustrate the credibility of our Mediator module in a more intuitively way,

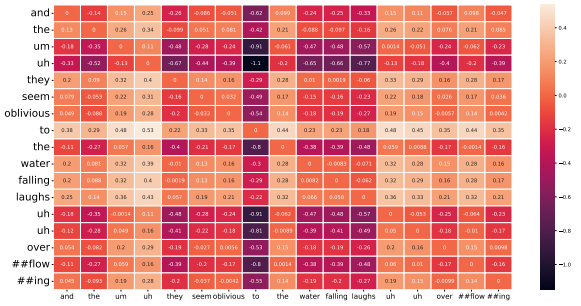


Figure 3: The perturbation degree heatmap of words from one test AD instance. The AD-specific features (*um*, *uh*) are assigned with lower perturbation degrees.

we conduct word-level case studies on the output of the Mediator module. As shown in Figure 3 and Figure 4, the heatmap plots illustrate relative perturbation differences between pairs of words from one test AD instance, where each cell's value represents the relative change in perturbation degree of one word compared to another. In each heatmap cell, the value $H(i, j)$ is calculated and normalized as follows:

$$H(i, j) = \frac{P_i - P_j}{P_i}. \tag{9}$$

The heatmap plots suggest that words like *um*, *uh* and repeatedly used personal pronoun like *i*, typically correlated to AD, show lower perturbation degrees relative to other words. The cells related to these words predominantly display values at or below zero, indicating minimal deviation from their original states. This pattern implies that the model has learned to recognize and preserve AD-specific features, thereby applying lower perturbations and minimizing noise for these critical words in AD detection scenarios.

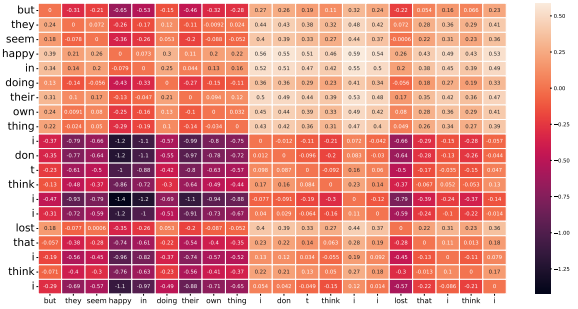


Figure 4: The perturbation degree heatmap of words from one test AD instance. The AD-specific features (repeatedly used pronoun: *i*) are assigned with lower perturbation degrees.

Class	N	P	$\Delta P(\%)$	Avg.	ΔAvg.(%)	$M\uparrow$	$M\downarrow$
AD	24	44.83	-34.33	1.87	-34.23	0.18	0.04
Non-AD	24	60.22	25.56	2.51	25.50	0.30	0.05

Table 6: Comparative analysis of perturbation degrees (PD) between AD and Non-AD test instances. This table summarizes the number of instances (N), the overall PD (P), the relative PD difference ($\Delta P(\%)$), the average PD (Avg.), the relative mean difference in PD (ΔAvg.(%)), and the range of PD, with maximum ($M\uparrow$) and minimum ($M\downarrow$) values for each class.

Considering the presence of AD-correlated features in Non-AD samples, relying solely on AD-specific features for detection does not align with the expected efficacy of a robust and sensitive AD discriminative model. In the AFP task, although the Mediator module may effectively detect AD by controlling PD of AD-specific features, it remains unclear whether this detection only depends on these features while potentially neglecting the contextual information in texts. To ensure model robustness and sensitivity, the Mediator module should assign higher perturbation degrees to Non-AD instances compared to AD instances. Simultaneously, it needs to astutely discern various linguistic patterns within instances and apply appropriate perturbation degrees.

Scenario 2: Comparison between AD and Non-AD (CAN) Here, we perform a comparison analysis to underscore the capability of our proposed method in differentiating AD patterns. Table 6 presents a comparative analysis of Perturbation Degrees (PD) for AD and Non-AD test instances. The AD class shows an overall PD of 44.83, with a relative PD difference of -34.33%. The average PD for AD instances is 1.87, with a relative mean difference of -34.23%, which effectively affirms the Mediator module's discriminative capability.

Additionally, the PD range for AD instances is narrow, with a maximum ($M\uparrow$) of 0.18 and a minimum ($M\downarrow$) of 0.04, indicating a fine-grained application of perturbations. Conversely, the Non-AD instances show an overall PD of 60.22, with a relative increase of 25.56%. The average PD is 2.51, with a relative mean difference of 25.50%. The broader PD range observed in Non-AD instances, with a maximum of 0.30 and a minimum of 0.05, indicates a higher degree of variability in perturbation application. This analysis verifies the Mediator module's ability in controlling perturbation degrees with the awareness of distinct linguistic patterns of AD and Non-AD instances.

Figures 2(b) and 2(c) further visually emphasize the above findings. AD instances in the test set reveal a clear trend of the Mediator module retaining clinically important information with minimal perturbation. As shown in Figure 2(b), object, location and action words, which are clinically important in AD detection, exhibit a higher degree of retention, whereas other features with less significance (e.g., pronouns and subjects) are subject to a higher degree of perturbation. In contrast, Figure 2(c) demonstrates more uniform perturbations for all word types for non-AD instances, indicating a lower specificity of feature retention.

To summarize, our Mediator module exhibits high credibility in effectively identifying and retaining linguistic features that are highly relevant to AD. Specifically, the Mediator module can effectively control the degree of perturbation for different types of words; it applies higher perturbations to words with lower relevance to clinically important features, but assigns lower perturbations to clinically important features. Moreover, the Mediator module not only focuses on the presence of clinically important features, but also takes into consideration their linguistic characteristics and contextual information within sentences when determining an appropriate degree of perturbations needed. The experimental results affirm the credibility of our Mediator module and emphasize the potential capability of our proposed method as a reliable tool in clinical diagnosis.

5 Discussion and Conclusion

In this paper, we propose a simple yet effective data augmentation framework (SPZ) for Alzheimer's disease detection under data-limited scenarios.

SPZ comprises two key components: the Generator module that generates token-level perturbations following skewed distributions, and the Mediator module that controls the perturbation degree for each word. The augmented textual representations are ultimately derived via zonal-mixing, enabling accurate AD detection.

Our SPZ method fundamentally differs from GAN-based DA methods that primarily focus on generating entirely new samples through training generative networks and often require large amounts of training data. In contrast, our SPZ method applies controlled semantic perturbations to existing text samples, with the goal of enhancing textual representations for AD detection. This approach preserves the linguistic integrity of the original data while enhancing the model's sensitivity to AD-specific linguistic patterns. Through generating controlled perturbations, SPZ offers the advantage of pinpointing subtle linguistic features associated with AD, thus yielding superior AD detection performance especially under data-limited settings.

Experimental results and detailed analyses demonstrate the potential of introducing semantic perturbations into textual representations to improve the accuracy of AD detection. Compared to a series of strong and competitive baselines, our SPZ method with controlled perturbations achieves the best results on the ADReSS challenge dataset. Additionally, the findings in extended analysis provide a basis for further exploring explicit feature inference and investigating credible AD detection.

Limitations and Future Work

Our SPZ method has demonstrated competitive performance on the ADReSS challenge dataset, but several potential limitations warrant further consideration. First, in our attempt to achieve a high recall to maximize the identification of potential AD cases, we observe a slight decrease in precision. Therefore, achieving a balanced tradeoff between precision and recall requires further investigation. Second, despite the data scarcity challenge in AD detection, the effectiveness of our SPZ method requires further evaluation on additional datasets. Third, although designed as a plug-and-play solution, the potential applicability of our SPZ method to other text classification tasks under low-resource scenarios requires further scrutiny.

In our future work, we will further explore new strategies for achieving a better balance precision and recall in AD detection. We plan to conduct a thorough analysis to determine which perturbations can most effectively enhance the model's sensitivity to AD-specific linguistic features. This investigation will refine our methodology for improved effectiveness, which is crucial for extending the model's utility as a reliable screening tool in clinical and real-world settings. In addition, we will further validate the efficacy of our SPZ method across diverse AD detection datasets, and explore its potential for application in other text classification tasks under low-resource scenarios.

Ethics Statement

The dataset used in this research originates from the ADReSS challenge, with strict protocols for minimal personal information and safeguarded against unauthorized access, adhering to ethical guidelines. The study's focus on English transcript text data limits the representation of Alzheimer's disease diagnoses to a specific cultural context. Designed for academic purposes, the model's diagnostic applicability in real-world clinical settings is limited, posing potential risks if deployed outside research environments.

Acknowledgements

This research was supported by the Open Project of Xiangjiang Laboratory [No.22XJ03004, No.22XJ03009], National Natural Science Foundation of China [62172449,72374070], Hunan Provincial Natural Science Foundation of China [2022JJ3021], Training Program for Excellent Young Innovators of Changsha [kq2107004], The science and technology innovation Program of Hunan Province [2022RC1105] and Industry-University-Research Innovation Fund of Chinese University[2021ITA01023]. This work was supported in part by the High Performance Computing Center of Central South University.

References

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

Hongmin Cai, Xiaoke Huang, Zhengliang Liu, Wenxiong Liao, Haixing Dai, Zihao Wu, Dajiang Zhu, Hui Ren, Quanzheng Li, Tianming Liu, et al. 2023.

Exploring multimodal approaches for alzheimer's disease detection using patient speech transcript and audio data. *ArXiv preprint*, abs/2307.02514.

Edward L Campbell, Laura Docío Fernández, Javier Jiménez Raboso, and Carmen García-Mateo. 2021. Alzheimer's dementia detection from audio and language modalities in spontaneous speech. In *IberSPEECH*.

Xuchu Chen, Yu Pu, Jinpeng Li, and Wei-Qiang Zhang. 2023. Cross-lingual alzheimer's disease detection based on paralinguistic and pre-trained features. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE.

Michael A. Deture and Dennis W. Dickson. 2019. The neuropathological diagnosis of alzheimer's disease. *Molecular Neurodegeneration*, 14(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junwen Duan, Fangyuan Wei, Jin Liu, Hongdong Li, Tianming Liu, and Jianxin Wang. 2023. CDA: A contrastive data augmentation method for alzheimer's disease detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1819–1826.

Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408.

Szatloczki Greta, Hoffmann Ildiko, Vincze Veronika, Kalman Janos, and Pakaski Magdolna. 2015. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in Aging Neuroscience*, 7.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3:642517.

Zhiqiang Guo, Zhaoci Liu, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. 2020. Text classification by contrastive learning and cross-lingual data augmentation for Alzheimer's disease detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6161–6171, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anna Hlédiková, Dominika Woszczyk, Alican Akman, Soteris Demetriou, and Björn Schuller. 2022. Data augmentation for dementia detection in spoken language. *ArXiv preprint*, abs/2206.12879.

Anna Hlédiková, Dominika Woszczyk, Alican Akman, Soteris Demetriou, and Björn Schuller. 2022. Data augmentation for dementia detection in spoken language.

Loukas Ilias and Dimitris Askounis. 2023. Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. *Knowledge-Based Systems*, 277:110834.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. 2020. Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2217–2221. ISCA.

Ning Liu, Kexue Luo, Zhenming Yuan, and Yan Chen. 2022. A transfer learning method for detecting alzheimer's disease based on speech and natural language processing. *Frontiers in Public Health*, 10:772592.

Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. 2021. Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7323–7327. IEEE.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2172–2176. ISCA.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Amit Meghanani, C. S. Anoop, and Angarai Ganesan Ramakrishnan. 2021. Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fasttext and cnn models. *Frontiers in Computer Science*, 3.

Jekaterina Novikova. 2021. Robustness and sensitivity of BERT models predicting Alzheimer's disease from text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 334–339, Online. Association for Computational Linguistics.

Gokul Prabhakaran, Rajbir Bakshi, et al. 2018. Analysis of structure and cost in a longitudinal study of alzheimer's disease. *Journal of Health Care Finance*, 44(3).

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. 2021. Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21:1–14.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. 2020. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2212–2216. ISCA.

Jianlin Su. 2020. bert4keras. `https://bert4keras.spaces.ac.cn`.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020a. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2162–2166. ISCA.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020b. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2162–2166. ISCA.

Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A. Batsis, and Robert M. Roth. 2021. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3790–3794. ISCA.