Data Augmentation for Dementia Detection in Spoken Language

Anna Hlédiková[1], Dominika Woszczyk[1*†], Alican Acman[1], Soteris Demetriou[1], Björn Schuller[1;2]*

[1]Department of Computing, Imperial College London, London, UK
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

anna.hledikova20@imperial.ac.uk, d.woszczyk19@imperial.ac.uk, a.akman21@imperial.ac.uk,
s.demetriou@imperial.ac.uk, bjoern.schuller@imperial.ac.uk

Abstract

Dementia is a growing problem as our society ages, and detection methods are often invasive and expensive. Recent deep-learning techniques can offer a faster diagnosis and have shown promising results. However, they require large amounts of labelled data which is not easily available for the task of dementia detection. One effective solution to sparse data problems is data augmentation, though the exact methods need to be selected carefully. To date, there has been no empirical study of data augmentation on Alzheimer's disease (AD) datasets for NLP and speech processing. In this work, we investigate data augmentation techniques for the task of AD detection and perform an empirical evaluation of the different approaches on two kinds of models for both the text and audio domains. We use a transformer-based model for both domains, and SVM and Random Forest models for the text and audio domains, respectively. We generate additional samples using traditional as well as deep learning based methods and show that data augmentation improves performance for both the text- and audio-based models and that such results are comparable to state-of-the-art results on the popular ADReSS set, with carefully crafted architectures and features[1].

Index Terms: dementia detection, data augmentation, speech.

1. Introduction

Dementia is the common term to describe a decline of cognitive abilities, such as memory, problem-solving or language that can severely impact an individual's communication abilities and everyday life in general. This condition, most often caused by Alzheimer's disease (AD), affects close to 60 million people and is the 7th leading cause of death, globally [1]. Fortunately, there has been a great focus across different fields not only to develop treatment for the condition but also to detect it, as an early diagnosis is key to help individuals to plan their future.

Recent challenges such as ADReSS and ADReSSo brought the focus to using speech and/or transcripts for AD detection [2, 3]. Detecting dementia, especially in its early stages, is a complex task which requires identifying subtle changes to semantics, vocabulary, or sentence-level structure. Information from audio data can provide beneficial information, e. g., about pauses or speech rate [4]. Subsequent work explored large pre-trained models and various DNN embeddings to extract fine-grained representations for text and audio features and achieved good results on the challenge held-out test set, considering the limited data available [5, 6, 7, 4, 8, 9]. Linguistic approaches have shown to be the most discriminative, even with simple models such as SVM. However, current state-of-the-art models implement hand picked features and tailored feature engineering techniques [10].

---

*Authors contributed equally.
†Work done outside of Amazon.
[1]The code is available at https://github.com/hl-anna/DA4AD

Audio classifiers benefit greatly from pre-trained embeddings such as the one introduced by [11], which have the advantage of being automatically extracted, compared to designed feature sets such as ComParE or eGeMAPS [12].

Collecting quality data on AD, be it transcripts or audio samples, can be difficult. Patients are hard to access, have to be evaluated for their mental state and the sessions have to be properly recorded and transcribed. Machine learning with limited data is a common challenge across tasks and domains, as limited amount of training data hinders the models' ability to generalise well and train properly. Some approaches addressing this problem explore different architectures and ensemble models [13, 8, 14, 15], feature engineering [16, 17], and data augmentation techniques [18, 19, 20, 21]. In this work, we focus on the latter. Data augmentation has been quite successful in image processing tasks, as one can perform many transformations without changing the object and label of the images. Although some of those techniques can be transferred to speech processing, they need to be carefully chosen for the task at hand and take into account the time dimension, too. Augmenting text data is more challenging, as it is highly sensitive to changes and one must be careful not to affect key label-specific characteristics. Nevertheless, several techniques have been explored by [22] and [23] in surveys which cover various tasks related to NLP or speech processing, extensively. Given the multi-modal nature of AD detection, suitable augmentation strategies are required for both audio and text. Individual text augmentation techniques have been applied in AD detection work [18, 21, 20], but no work offers an overview of augmentation strategies.

In this paper, we aim to expand the scope of previous surveys on data augmentation and evaluate label preserving data augmentation techniques on the task of Alzheimer's disease detection for both NLP and speech processing. We examine a total of 17 approaches (7 for text, 10 for audio) and evaluate them on baseline and SOTA models. We work with the two domains separately, as well as with a fusion of both. Our experiments show that the selected data augmentation methods were able to preserve the class labels and that augmentation itself is effective in improving the model's generalisation abilities on the ADReSS dataset. This has been shown using neural-based and more traditional models, on both the text and audio domains. We believe that this work can contribute to the field and offer an inexpensive way to improve AD detection methods in the future. Our contributions are as follows:

1. We empirically evaluate text and speech data augmentation techniques for the task of dementia detection.

2. We analyse the level of label preservation for each augmentation.

3. We achieve performance on par with state-of-the-art model for both text and audio domains via data augmentation.

2. Data Augmentation Strategies

2.1. Text Augmentations

2.1.1. Noise

Noise can be introduced by means of random deletions, insertions, or substitution at character, word, or sentence level. For the text domain, we employ three noise injection methods: sentence deletion (SD), easy data augmentation (EDA) [24], and Mixup [25]. For the SD approach, we delete one or more randomly selected sentences from each transcript. We found that removing one to four sentences (at random) yields the best results, so the average SD operation reduces a transcript by 10 %. The Mixup approach splits a transcript into two halves and swaps these with a different sample within the same class. EDA randomly replaces, inserts, or deletes words within a document. This approach can affect the semantics or introduce grammar mistakes. We follow the EDA implementation from the original paper and set $\alpha = 0.05$.

2.1.2. Lexical substitution

An intuitive approach to augmenting data while preserving the label is to substitute words with their synonyms. A simple option is to replace words with their WordNet [26] synonyms or hyperonyms from the Knowledge Graph. The words to substitute can be sampled at random, using TFIDF-weights or by analysing the rate of change when the word is deleted. However, this is a tricky operation in case of AD detection. Words might be substituted by alternatives causing the sample to loose its discriminative features, for example by going from familiar language to elevated vocabulary. Some works use pre-trained embeddings and look for close words given the feature space, using geometric distances. These approaches suffer from the same issues. To tackle this limitation, one can use contextual augmentation, i. e., use language models to replace words given a sentence and learnt priors for each label [27]. In this work, we implemented a version of contextual augmentation using the NLPAug library [28], based on the RoBERTa model. We set $p = 0.1$ and $top\_k = 20$, i. e., replacing 10 % of words by picking randomly for the top 20 alternatives.

2.1.3. Paraphrase

Paraphrasing consists in rewording or changing the structure of a document while maintaining its original meaning. Earlier approaches include statistical and rule-based models or selecting sub-sentences from a pool of stored templates. More recently, neural models have been adopted to generate paraphrases directly in an end-to-end fashion. A sub-task of paraphrasing is back-translation (BT), which creates new samples by translating text from one language to another and back to the original one, thus exploiting differences in structure and vocabulary across languages. Similarly, text summarisation aims to restructure sentences in a concise way. This approach is less desirable diversity-wise, as the target space is more limited. We split each document on sentence level and run the pre-trained paraphrase model Pegasus [29], a model with 223M parametres, to generate paraphrases of length $n = 60$ tokens and with temperature $t = 1$.

2.1.4. Text Generation

Generative models are a popular approach to data augmentation for text. Rule-based text generation methods have now been supplanted by DNNs, sequence-to-sequence, and transformer-based methods with large models trained on even larger datasets. One such model is OpenAI's GPT-2 [30], which we use to perform text generation. GPT-2 is pretrained on the WebText corpus – a large collection of internet blogs, posts, and pages, but can be finetuned with a relatively small set of samples to adapt the feature distribution. We use Hugging'face's[2] model with 117M parametres. The model is finetuned on the entire training set, conditioned with the sample label class. The label is then used as a prompt for the model at inference time.

2.2. Audio Augmentations

2.2.1. Standard transformations

Standard transformations can be defined as a function g applied on the sample x at time step t such that $x'(t) = g(x(t))$. We apply these transformations via the `nlpaug`[3] python package.
Noise addition is one of the simplest transformations. The noise ϵ is usually sampled from a Gaussian distribution $N(\mu, \sigma)$, in our case $N(0,1) * 0.002$, and is applied at step t either in the time or frequency domain. The new sample can be defined as $x'(t) = x(t) + \epsilon$.
Time stretching changes the speed of a signal by a factor of α, without changing the pitch. For $\alpha < 1$, the signal is slowed down and lengthened and sped up for $\alpha > 1$.
Pitch shifting perturbs the pitch by n fractions of an octave. The signal is passed through a Short-Term Fourier Transform (STFT) to obtain the spectrogram in the frequency domain to apply the change to the pitch, without changing the length or speed of the signal. We adjust the pitch by a scale factor in $(-10, 10)$.
Time shift adds padding on the left or the right of a signal. We shift time of the whole audio by a duration of 0.5 seconds.
Loudness shift increases or decreases the magnitude of a signal x by β DB or by scaling the signal by a factor α s. t. $x'(t) = \alpha x(t)$, where the volume is decreased for $\alpha < 1$. We adjust loudness of each sample by a scale factor within $(0.3, 3)$.
Normalisation of the signal can be performed given minimum and maximum amplitudes or by standardising it. We simply normalise the signal by dividing it by the maximum amplitude.
Time masking consists in masking consecutive time bands on a log Mel-spectrogram, where $t \,\epsilon [t_0, t_0 + t]$ is sampled from a uniform distribution $U(0, T)$, $t_0 \sim U(0, T - t)$, and T is the maximum length of masking steps. We apply masking operation with a coverage of 0.3.
Frequency masking applies a mask on consecutive frequency bands on a log Mel-spectrogram, where $f \sim U(f_0, f_0 + f)$ and $f_0 \sim [0, V - f]$, where V is the maximum number of Mel-frequency channels. We implement two strategies that combine several transformations: *Random* and *SpecAugment*. The *Random* strategy selects a transformation for each sample at random and *SpecAugment* combines time warping, frequency, and time masking on the log Mel-spectrograms [31].

2.2.2. Vocal Tract Length Perturbation (VTLP)

Vocal Tract Length Perturbation (VTLP) operates on the log-Mel-spectrogram itself [32]. Similarly to time warping, VTLP performs warping on a frequency f. We apply a VTLP operation on the whole audio with a scaling factor of $(0.5, 3)$.

---

[2]https://huggingface.co/gpt2
[3]https://github.com/makcedward/nlpaug

2.2.3. Generative Models

Deep neural networks are now dominating generative models such as auto-encoders, generative adversarial networks (GANs) or sequence-to-sequence models such as LSTM networks. We perform zero-shot voice conversion using a FragmentVC [33] pre-trained model[4]. It performs better at speaker identity conversion than popular state-of-the art AutoVC [34] while preserving prosody and durations, which is actually desirable in our case to preserve class-specific style.

3. Experiments & Results

3.1. Datasets

3.1.1. ADReSS and Dementiabank

We work with two datasets, the ADReSS dataset [2][5] and Dementiabank.[6]. Both consist of recordings of a Cookie theft picture description task from individuals diagnosed with various stages of dementia. The datasets also contain manual transcriptions of the samples with dysfluency annotations, and the ADReSS set is a subset of Dementiabank. The speakers in the ADReSS set are American and have been selected to balance for gender and age, and they split evenly into control (HC) and dementia (AD) groups (54 train and 24 test for each class, total of 156 samples). Dementiabank has three times as many samples, but is imbalanced, lacks a fixed test set and some samples contain a lot of noise. The main focus of this paper is thus on the ADReSS set, which is of better quality and allows for an easy comparison of results with other works. However, we also examine the effect of augmentation on a larger dataset such as Dementiabank.

3.1.2. Preprocessing

Transcripts from both datasets were annotated using the CHAT coding system which includes POS tags and grammatical dependencies but also pauses and other dysfluencies. We read each CHAT file using the python library `pylangacq` [7]. We retain the cleaned sentences uttered by the participant, with no additional dysfluency annotation. For the audio recordings, we use the normalised, noise reduced wavefiles sampled at 44 kHz. We split them into chunks of $10s$ with a stride of $2s$ and apply the data augmentation either on the signal or spectrograms computed using the `librosa` [8] Python library.

3.2. Models

3.2.1. Text-based detection models

To perform the linguistic classification task, we reimplement [4, 6] using a pre-trained BERT model on the transcripts. We use the base size model (12 layers) with the following patametres: learning rate of 2e-5, 8 epochs, and input length limit of 256 tokens. We feed the whole document as a datapoint, keeping the dysfluencies annotations. To focus on text only, we ignore the pauses extracted for the audio samples. The BERT model is widely used in the field of NLP and the authors from [4] have achieved SOTA results on the ADReSS Challenge 2020. As a more traditional approach, we also implement an SVM with an

RBF kernel with $gamma = 0.01$ and $C = 1$. We use TF-IDF features as input.

3.2.2. Acoustic-based detection models

The audio classifier we use in our experiments is the Audio Spectrogram Transformer (AST) [11], which adapts the transformer architecture for audio. We extract Mel spectrograms as input to the AST model. We tune the model for 2 epochs, using a learning rate of 1e-6 and the Adam optimiser. Similarly as for text, we also implement a simpler baseline which is a Random Forest (RF) model. We extract eGeMAPS [12] sets of features to train the RF model, which has been shown to be a better optimised set of features compared to ComParE on the DementiaBank dataset [2]. The RF model is trained with 100 trees.

3.2.3. Fusion

Our fusion model consists of the three best performing augmentation approaches at cross-validation time. We perform a majority vote to predict the final label and report the result in Table 3.

3.3. Experiment Setup

All models are trained on one NVIDIA Tesla K80, 16 GB RAM; all code is written in Pytorch 1.10.0+cu111. For all models, we report the accuracy as well as the sensitivity. Given the nature of the task, we believe it is important to put emphasis on the ability of models to detect a case of AD rather than just the accuracy. We perform 10-fold and 5-fold cross validation for text and audio, respectively, and report the average accuracy across 5 different seeds. Implementation details are available in our repository. We further evaluate the augmentation approaches in terms of label preservation using a model trained only on the original data. We then measure the accuracy for both classes. We also compute the KL-divergence on extracted features and the Mel-spectrogram distortion between the original and augmented samples for audio. For text, we compute the Levenshtein distance, semantic distance, and type token ratio (TTR). For semantics, we extract sentence embeddings with the *Huggingface* model *all-mpnet-base-v2*[9] and compute the cosine distance.

3.4. Results

The results of our experiments evaluated on the ADReSS dataset are displayed in Tables 1 and 2 for text and audio, respectively, and a comparison to the challenge baseline as well as state-of-the-art results can be found in Table 3. We can see that while DNN-based approaches to augmentation are an effective way to improve model performance, some traditional and less computationally expensive methods work comparably well. For the text augmentation, paraphrasing methods showed the best results, whether it is by *Back-translation* (BT), which achieves the highest score of 84% for Russian (RU) and 85% for German (DE) for BERT and SVM respectively, or by using the Pegasus model. However, simpler noise injection methods such as *EDA* and *Mixup* are not far behind in terms of the performance increase. Generally, the augmentations reduced overfitting for the BERT Model while the SVM tends to overfit for almost all techniques. However, we can observe that the introduction of new samples improved both models' generalisation abilities via *Text generation* for text and with *FragmentVC* for audio, with second best results for AST (73%) and best for RF (69%).

---

[4]https://github.com/yistLin/FragmentVC

[5]The dataset is accessible upon request at http://www.homepages.ed.ac.uk/sluzfil/ADReSS/ and is governed by the Creative Commons CC BY-NC-SA 3.0 license.

[6]https://dementia.talkbank.org/

[7]https://pylangacq.org/

[8]https://librosa.org/doc/latest/index.html

[9]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Similarly as for text, while individual traditional methods only offered improvement in some cases for the audio domain, the approach of *Random* ensemble performed just as well as a more complex DNN-based method of voice conversion. We can also see that the results of our experiments are often considerably higher for the test set than for cross-validation. While this might partly be simply due to the small size of the dataset we work with, it also shows the role of data augmentation as a regularisation method. We also note the fusion of the top-3 best models in Table 3 and achieve the best results with 86%.

We also evaluate the techniques on the DementiaBank, but we did not notice considerable improvement on the model's performance. We attribute this to the more noisy nature of the dataset and its larger size.

		BERT		SVM		BERT	SVM
Method		CV	Test	CV	Test	Test Sensitivity	
None		.80 (.04)	.78 (.06)	.83 (.00)	.81 (.00)	.81(.04)	.71 (.00)
SD		.74 (.01)	.77 (.02)	.83 (.00)	.81 (.00)	.68 (.06)	.75 (.00)
Mixup		.79 (.04)	.82 (.04)	.85 (.00)	.75 (.00)	.79 (.06)	.67 (.00)
EDA		.80 (.02)	.83 (.03)	.84 (.00)	.79 (.00)	.81 (.04)	.71 (.00)
Paraphrasing		.82 (.05)	.81 (.03)	.84 (.00)	.81 (.00)	.78 (.04)	.75 (.00)
Text Gen.		.71 (.02)	.81 (.05)	.77 (.00)	.81 (.00)	**.84 (.07)**	.75 (.00)
Context Aug.		.80 (.05)	.80 (.05)	.83 (.00)	.77 (.00)	.80 (.06)	.67 (.00)
BT (DE)		**.82 (.02)**	**.84 (.02)**	**.86 (.00)**	.81 (.00)	.81 (.06)	.75 (.00)
BT (RU)		.80 (.03)	.82 (.01)	**.86 (.00)**	**.85 (.00)**	.83 (.06)	**.79 (.00)**

Table 1: *Accuracy and test Sensitivity results for Text augmentations on the ADReSS set, with standard deviation in parentheses. For the SVM model, once parameters are fixed, changing the random seed does not affect results, hence the standard deviation is zero. SD = random sentence deletion, BT = back-translation from either German (DE) or Russian (RU). Best results for each model are highlighted.*

		AST		RF		AST	RF
Method		CV	Test	CV	Test	Test Sensitivity	
None		.62 (.13)	.70 (.03)	.58 (.14)	.67 (.02)	1.00 (.00)	.79 (.00)
Random		.63 (.09)	**.74 (.03)**	.79 (.12)	.68 (.02)	.92 (.08)	.79 (.00)
Masking		.56 (.11)	.70 (.04)	.69 (.16)	.67 (.02)	**.98 (.02)**	.75 (.00)
Loudness		**.64 (.07)**	.70 (.01)	.78 (.14)	**.69 (.02)**	.83 (.13)	.79 (.00)
Noise		.62 (.17)	.71 (.04)	.69 (.15)	.64 (.01)	.83 (.04)	.79 (.00)
Pitch		.61 (.13)	.72 (.02)	.68 (.15)	.64 (.02)	.83 (.00)	**.88 (.00)**
VTLP		.61 (.07)	.70 (.04)	.79 (.13)	.67 (.02)	.79 (.04)	.75 (.00)
Shift		.59 (.11)	.72 (.02)	.79 (.14)	.66 (.02)	.79 (.04)	.75 (.00)
Speed		.63 (.07)	.70 (.02)	**.80 (.10)**	.64 (.02)	.81 (.02)	.75 (.00)
Normalisation		.56 (.07)	.72 (.03)	.76 (.11)	.68 (.01)	.79 (.04)	.83 (.00)
SpecAugment		.56 (.10)	.70 (.02)	.53 (.07)	.67 (.02)	.90 (.02)	.79 (.00)
FragmentVC		.62 (.04)	.73 (.04)	.54 (.04)	**.69 (.03)**	.94 (.06)	**.88 (.00)**

Table 2: *Accuracy and test Sensitivity results for Audio augmentations on the ADReSS set, with standard deviation in parentheses. The models used are a Random Forest (RF) and an Audio Spectrogram Transformer (AST). VTLP = Vocal tract length perturbation. The best results for each model are highlighted.*

We note the label preservation performance as well as divergence from the original data distribution methods in Table 4. Methods with the highest label preservation scores tend hurt rather than benefit the model performance, suggesting that in those cases the augmented samples remain too similar to the original ones and the model ends up overfitting. At the same time, methods that diverge too much are likely to change key discriminative features of AD, also leading to poor performance. We can see that while noise injection and generative methods for audio have a considerably higher divergence score than the other traditional methods, they still have a beneficial effect on training. Comparing Table 1 and Table 4, we can see that although DNN-based approaches lower the label preservation, the perfor-

	Model	Features	Accuracy
Baseline	LDA* [2]	Text	.75
	LDA* (ComParE features) [2]	Audio	.63
SOTA	ERNIE (ensemble) [4]	Text	**.85**
	Music-Linear-BOW [8]	Audio	**.74**
	Fusion [8]	Audio + Text	**.90**
Our models	SVM (BT RU)	Text	**.85**
	AST (Random)	Audio	**.74**
	Fusion: AST (Random) + SVM (BT RU) + BERT (BT DE)	Audio + Text	.86

Table 3: *Comparison between our systems and SOTA on the ADReSS dataset. There are no baseline models or easily comparable work for the Dementiabank dataset. For our models, the model name and augmentation method used is provided. The best result is highlighted for each domain.*
LDA = Linear discriminant analysis.

mance on the test set and cross-validation sets indicate that the introduced noise decrease the models overfitting. In particular, *Text generation* is the most disruptive in terms of all metrics but in fact improves on the accuracies and test sensitivity for both BERT (81% & 84%) and SVM (81% & 75%). *Backtranslation* seems to strike the right balance between deviating from the training set (semantics, introduced new tokens) and preserving meaningful features, and similarly *FragmentVC* for audio.

Domain	Method	Label Pres.		Δ		
		Acc.	F1	Levenshtein	Semantics	TTR
Text	**None**	.97	1.00	0	.00	.00
	SD	.90	.90	71	.08	.03
	Mixup	.92	.92	258	.20	.09
	EDA	.88	.88	40	.07	.01
	Paraphrasing	.79	.75	267	.19	.10
	Text Generation	.62	.69	411	.39	.14
	Context Aug.	.88	.89	48	.09	.03
	BT (DE)	.83	.83	178	.15	.05
	BT (RU)	.83	.84	228	.21	.03
				Divergence	Mel. Distortion	
Audio	**None**	1.00	1.00	.00	.00	-
	Random	.90	.91	.07	5.4	-
	Masking	.87	.90	.07	11.7	-
	Loudness	.95	.96	.07	1.2	-
	Noise	.84	.85	.20	6.9	-
	Pitch shift	.80	.83	.07	3.5	-
	VTLP	.98	.98	.06	5.2	-
	Time shift	.99	.99	.05	5.9	-
	Speed	.95	.96	.06	5.1	-
	Normalisation	.88	.90	.06	3.8	-
	SpecAugment	.58	.60	.37	19.6	-
	FragmentVC	.58	.52	.39	9.7	-

Table 4: *Label preservation and descriptors distances for approaches on Text and Audio domains on the ADReSS set.*

4. Conclusion

In this paper, we investigated a range of data augmentation techniques for the domains of text and speech. We evaluate the strategies on the ADReSS challenge dataset for the task of dementia detection for neural, SVM, and Random Forest models as well as their late fusion. Additionally, we quantify the label preservation properties for all augmentation approaches. Our experiments show that most of the techniques can help the models generalise and reduce overfitting on the training set. Generally we observe that neural-based augmentations do not necessarily outperform simpler approaches such as adding noise or performing random deletions and substitutions. Our models achieve results comparable to state-of-the-art approaches on the ADReSS set by simply augmenting the training set. Their optimal combination needs to be analysed next.

5. References

[1] . Alzheimer's Disease International, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019," *The Lancet Public Health*, vol. ISSN: 2468-2667, 2022, doi.org/10.1016/S2468-2667(21)00249-8.

[2] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: the adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[3] ——, "Detecting cognitive decline using speech only: The ADReSSo challenge," 2021.

[4] 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," pp. 2162–2166, 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2516

[5] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in Computer Science*, vol. 3, p. 22, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomp.2021.624683

[6] 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," 2020.

[7] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. Nallanthighal, H. Christensen, D. Blackburn, B. Schuller, M. Magimai-Doss, H. Strik, and A. Härmä, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," 10 2020.

[8] 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*, vol. 9, pp. 88 377–88 390, 2021.

[9] J. Glass *et al.*, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, p. 3833, 2021.

[10] 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*, vol. 13, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2021.642647

[11] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[13] L. Ilias and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in Aging Neuroscience*, p. 205.

[14] Y. Santander-Cruz, S. Salazar-Colores, W. J. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, "Semantic feature extraction using sbert for dementia detection," *Brain Sciences*, vol. 12, no. 2, p. 270, 2022.

[15] N. An, H. Ding, J. Yang, R. Au, and T. F. Ang, "Deep ensemble learning for alzheimer's disease classification," *Journal of biomedical informatics*, vol. 105, p. 103411, 2020.

[16] V. N. Garla and C. Brandt, "Ontology-guided feature engineering for clinical text classification," *Journal of biomedical informatics*, vol. 45, no. 5, pp. 992–998, 2012.

[17] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[18] A. Roshanzamir, H. Aghajan, and M. Soleymani Baghshah, "Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech," *BMC Medical Informatics and Decision Making*, 2021.

[19] Y. Guo, C. Li, C. Roan, S. Pakhomov, and T. Cohen, "Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task," *Frontiers in Computer Science*, vol. 3, p. 26, 2021.

[20] Z. Ye, S. Hu, J. Li, X. Xie, M. Geng, J. Yu, J. Xu, B. Xue, S. Liu, X. Liu *et al.*, "Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6433–6437.

[21] Z. Liu, Z. Guo, Z. Ling, and Y. Li, "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)*. IEEE, 2021, pp. 7323–7327.

[22] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," *arXiv preprint arXiv:2105.03075*, 2021.

[23] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in nlp," *arXiv preprint arXiv:2106.07499*, 2021.

[24] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," 2019.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[26] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to wordnet: An on-line lexical database," *Communications of the ACM*, 1995.

[27] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," 2018.

[28] E. Ma, "Nlp augmentation," https://github.com/makcedward/nlpaug, 2019.

[29] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[32] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[33] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5939–5943.

[34] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.