applied sciences

MDPI

Article

The Impact of Pause and Filler Word Encoding on Dementia Detection with Contrastive Learning

Reza Soleimani [1], Shengjie Guo [1], Katarina L. Haley [2], Adam Jacks [2] and Edgar Lobaton [1,*]

[1] Department of Electrical and Computer Engineering, North Carolina State University, Engineering Bldg II, 890 Oval Dr, Raleigh, NC 27606, USA; soleimani.reza1994@gmail.com (R.S.); sguo25@ncsu.edu (S.G.)
[2] Division of Speech and Hearing Sciences, Department of Allied Health Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27559, USA; katarina_haley@med.unc.edu (K.L.H.); adam_jacks@med.unc.edu (A.J.)
* Correspondence: ejlobato@ncsu.edu

Abstract: Dementia is primarily caused by neurodegenerative diseases like Alzheimer's disease (AD). It affects millions worldwide, making detection and monitoring crucial. This study focuses on the detection of dementia from speech transcripts of controls and dementia groups. We propose encoding in-text pauses and filler words (e.g., "uh" and "um") in text-based language models and thoroughly evaluating their impact on performance (e.g., accuracy). Additionally, we suggest using contrastive learning to improve performance in a multi-task framework. Our results demonstrate the effectiveness of our approaches in enhancing the model's performance, achieving 87% accuracy and an 86% f1-score. Compared to the state of the art, our approach has similar performance despite having significantly fewer parameters. This highlights the importance of pause and filler word encoding on the detection of dementia.

Keywords: dementia; contrastive learning; deep learning; text classification; LLMs; NLP

Citation: Soleimani, R.; Guo, S.; Haley, K.L.; Jacks, A.; Lobaton, E. The Impact of Pause and Filler Word Encoding on Dementia Detection with Contrastive Learning. *Appl. Sci.* **2024**, *14*, 8879. https://doi.org/10.3390/app14198879

Academic Editor: Alexander N. Pisarchik

Received: 7 June 2024
Revised: 10 August 2024
Accepted: 25 September 2024
Published: 2 October 2024

1. Introduction

Dementia is a progressive cognitive disorder caused by neurodegenerative diseases, with Alzheimer's disease (AD) being the most prevalent form [1]. AD accounts for a significant majority of dementia cases, affecting millions worldwide. Given the extensive impact of AD and the current lack of a cure, early detection of dementia is crucial. Detecting the disease in its early stages can allow for timely intervention, which can slow the progression of symptoms and provide better management options. Early diagnosis can greatly improve the quality of life for individuals living with dementia, enabling them and their families to plan and access support sooner [2].

In recent years, deep neural networks (DNNs) have shown considerable promise in the detection of dementia from speech transcripts. By utilizing DNNs, models can be enhanced with data features that facilitate the early detection of dementia. Researchers have employed audio recordings, textual data, and biomedical imaging to detect dementia. In this paper, we focus specifically on textual data to leverage large language models (LLMs). LLMs are pre-trained models trained on vast corpora of data from various topics. Some well-known LLMs include BERT and its variants [3,4], GPT-3 [5], and others. These models contain contextual information within the text, enabling the extraction of syntactic and semantic information. Due to this contextualized training, they excel in downstream tasks such as semantic analysis, question answering, and named-entity recognition [6–8].

In the literature, researchers have employed various approaches for the detection of dementia based on speech transcripts. In early attempts, researchers in [9–11] focused on using lexical and syntactic information to detect dementia. These approaches involved analyzing factors such as the frequency of nouns and verbs, common word usage, language fluency, and other related linguistic features. In recent studies, using transfer learning

through LLMs has become very popular due to their powerful ability to capture complex language patterns and enhance model performance [12–15]. The authors in [14] proposed a sentenced-based pipeline that integrates various augmentation techniques, pre-trained LLMs, and classifiers to perform classification tasks. They thoroughly evaluated each augmentation technique, explored a range of pre-trained models, and experimented with different classifiers, including CNNs and RNNs. Additionally, the authors implemented various voting mechanisms to refine their results, presenting a detailed analysis of each component's impact on the overall performance. Other studies have introduced innovative methods, such as contrastive learning (CL), which involves learning from the data themselves to separate the feature space for each class [16,17]. In [16], the authors propose a multimodal approach which uses both text and audio using graph neural networks (GNNs) with CL, and study the impact of both modalities. Also, the authors in [18,19] explore the use of Transformers and attention-based approaches combined with CNNs and RNNs. Due to the limited data available in this field, several authors have proposed various augmentation techniques to mitigate this issue [20–22]. In [21], the authors proposed a generative approach for data augmentation by using variational auto-encoders (VAEs). Instead of raw text or speech, they augmented the lingual and acoustic features. These augmentation methods aimed to increase the dataset size, enhancing the model's ability to generalize and perform better [23].

In [12,24], the authors proposed a technique where pause information is encoded within the text, a feature typically used in speech-based analysis. Pauses are introduced using special characters, enabling the language model to recognize these features, which has been shown to enhance model performance. In [12], the authors utilized the ADReSS dataset [25] for their experiments and employed temporal word alignment to implement their methodology. Building on this idea, we apply a similar concept to the Pitt Corpus Cookie Theft dataset [26].

In this paper, we propose a new methodology using LLMs for the detection of dementia from transcriptions of speech by encoding in-text pauses and filler words, and study the impact on the performance of contrastive learning. We hypothesize and validate that the inclusion of these encodings and modeling schemes lead to a significant impact on performance, starting at 57% accuracy for a baseline model and increasing to 87% by our final model. This paper is organized as follows: In Section 2, the main materials and methods used are explained, which include dataset preparation, in-text pause encoding, and the contrastive learning scheme. In Section 3, the results for the experiments are presented. In Section 4, our results are discussed. Finally, in Section 5, the paper is concluded.

2. Materials and Methods

This section provides the data preparation and modeling steps. The main objective of this paper is to enhance the performance of detection models through the use of in-text encoding and contrastive learning, which can be considered a multi-task learning scheme. As shown in [27–29], multi-task learning can be beneficial in improving model performance. Below, we summarize our contributions:

- Proposing in-text pause and other language features (uh/um) encoding.
- Thoroughly evaluating different pauses to provide insight on how they affect model performance.
- Proposing to use contrastive learning to improve performance.
- Combining in-text pause encoding and DualCL in a multi-task manner.

As mentioned in Section 1, our proposed approach is similar in context to [12,24]. In [12], the authors utilized temporal word alignment to embed the pause information within the text. We take a different approach in our modeling that does not involve temporal word alignment, which follows the directions in [24]. We explore alternative methods to encode pauses, their different combinations, and language information directly within the textual data; this has not been attempted before (please see Table 1). This approach

allows us to investigate the impact of pause information on model performance, potentially offering new insights into dementia detection through textual analysis.

Another approach of interest in this paper is contrastive learning (CL). Researchers have adopted contrastive methods to perform dementia detection, relying on the data themselves to improve the representation space while enhancing model performance [30]. In this paper, we employ a technique called dual contrastive learning (DualCL) [31], which has demonstrated strong performance with general textual data. In Section 3.3, we provide a detailed explanation of this methodology.

Table 1. Different in-text encoding schemes.

Input Type	Example
Original	*"sen1 (. . .) sen2 (.),sen3 , (..) . . . "*
E_{I_0,S_0,F_0,V_0}	*"sen1 sen2 sen3 . . . "*
E_{I_1,S_0,F_0,V_0}	*"sen1 Lo sen2 Sh sen3 Me . . . "*
E_{I_0,S_1,F_0,V_0}	*"sen1 sen2 sen3 . . . Lo Sh Me . . . "*
E_{I_0,S_0,F_1,V_0}	*"sen1, sen2, sen3 . . . , #Sh #Me Me #Lo Lo"*
E_{I_0,S_0,F_0,V_1}	*"sen1, sen2, sen3 . . . ,#Sh #Me #Lo"*
E_{I_1,S_1,F_0,V_0}	*"sen1 Lo sen2 Sh sen3 Me . . . , Lo Sh Me . . . "*
E_{I_1,S_0,F_1,V_0}	*"sen1 Lo sen2 Sh sen3 Me . . . , #Sh Sh #Me Me #Lo Lo . . . "*
E_{I_1,S_0,F_0,V_1}	*"sen1 Lo sen2 Sh sen3 Me . . . , #Sh #Me #Lo"*
E_{I_0,S_1,F_1,V_0}	*"sen1, sen2, sen3 . . . , Lo Sh Me . . . , #Sh Sh #Me Me #Lo Lo . . . "*
E_{I_0,S_1,F_0,V_1}	*"sen1, sen2, sen3 . . . , Lo Sh Me . . . , #Sh #Me #Lo"*
E_{I_0,S_0,F_1,V_1}	*"sen1, sen2, sen3 , . . . , #Sh Sh #Me Me #Lo Lo . . . , #Sh #Me #Lo"*
E_{I_1,S_1,F_1,V_0}	*"sen1 Lo sen2 Sh sen3 Me . . . , Lo Sh Me . . . , #Sh Sh #Me Me #Lo Lo . . . "*
E_{I_1,S_1,F_0,V_1}	*"sen1 Lo sen2 Sh sen3 Me . . . , Lo Sh Me . . . , #Sh #Me #Lo"*
E_{I_1,S_0,F_1,V_1}	*"sen1 Lo sen2 Sh sen3 Me . . . , #Sh Sh #Me Me #Lo Lo . . . , #Sh #Me #Lo"*
E_{I_0,S_1,F_1,V_1}	*"sen1, sen2, sen3 . . . , Lo Sh Me . . . , #Sh Sh #Me Me #Lo Lo . . . , #Sh #Me #Lo"*
E_{I_1,S_1,F_1,V_1}	*"sen1 Lo sen2 Sh sen3 Me . . . , Lo Sh Me . . . , #Sh Sh #Me Me #Lo Lo . . . ,#Sh #Me #Lo"*

2.1. Dataset Preparation

In this study, the Pitt Corpus Cookie Theft dataset [26] from DementiaBank is used. These transcripts are rich in detail, including the patients' demographic information like gender and age, as well as clinical data such as dementia severity. Additionally, they contain syntactic details to ensure language consistency, timestamps, and dialogues between researchers and participants. This dataset contains 243 and 305 recordings and CHAT style transcriptions for control and dementia groups, respectively. Throughout our experiments, we use the ground-truth transcriptions for our training and evaluation. For our analysis, we specifically exclude certain special characters found in the conversations, such as "[//]", "&-uh", and "&=laughs", among others. Each of these symbols has its meaning, which can be found in the DementiaBank documentation [26].

2.2. In-Text Pause Encoding

In this section, an in-text pause encoding scheme is explained. In this methodology, a relationship between pauses and dementia diagnosis within the textual information is explored. Lately, this topic has been of interest in the literature [12,24], in which the classification is achieved based on the frequency and duration of pauses within the speech. In our previous work, presented in [24], preliminary experiments were conducted on in-text encoding.. The method is fully explained in this section. The current work expands on [24] by thoroughly examining the effect of each pause encoding and adding new textual cues (filler words) to the analysis. Also, a combination of each encoding with CL is studied. In Section 3, all the results are presented in detail.

An important reason to use textual information over audio recordings is that while audio recordings are very rich in information, they are very complex and require significant

computational resources to be properly utilized [32,33]. Additionally, transcripts from standard tests provide an additional layer of privacy. The transcripts of the audio can be obtained once and they can be used for training and evaluation. The transcription can be performed by very powerful tools such as the Whisper [34] and Wav2vec [35] models.

As mentioned in Section 2.1, the Cookie Theft dataset is used in the experiments in our study. We follow the same methodology as was introduced in [24] to construct the in-text pause encodings. In the dataset, some special characters indicate different pause lengths. To be specific, the symbols "(.)", "(..)", and "(...)" represent short, medium, and long pauses, respectively. These pauses are measured in seconds or fractions of a second, depending on the type of the pause. For short, medium, and long pauses, the pause lengths are under 0.5 s, 0.5–2 s, and over 2 s, respectively [12]. We replace these symbols with "Sh", "Me", and "Lo", to be inserted within the text. Also, a vector that contains the frequency of each pause is used in our analysis.

Given a sample text of the form

"seg1 (...) seg2 (.) seg3 (..) ...",

we introduce the following encodings: In-place (I), end-sequence (S), frequency (F), and vector (V). Each of these combinations can be present or absent in the encoding, which results in 16 different combinations in total. The base text takes the form

"seg1 seg2 seg3 ...".

The baseline, where no encoding and vector information is included, is represented by B_0. The experiments showed that models performed better when they were provided with a secondary numerical vector input corresponding to the frequencies of the pauses in the form

[#Sh, #Me, #Lo],

where #Sh, #Me and #Lo represent the number of short, medium, and long pauses, respectively. It should be mentioned that all the combinations include the secondary vector input, which has its secondary model.

For example, if a text only contains the in-place encoding, it is represented as E_{I_1,S_0,F_0,V_0}. In this case, our sample text takes the form

"seg1 Lo seg2 So seg3 Me ...".

The end-sequence encoding attaches all the pauses in the order that they are happening in the text to the end of the text as follows:

"seg1 seg2 seg3 ... Lo Sh Me ...".

The third encoding, frequency encoding, creates a text with the count of each pause type attached to each pause type (e.g., "Sh"), and is attached to the end of the text as follows:

"seg1 seg2 seg3 ... #Sh Sh #Me Me #Lo Lo".

Lastly, the vector encoding is similar to the frequency encoding, but the pause type is not included in the encoding. It can be shown as follows:

"seg1 seg2 seg3 ... #Sh #Me #Lo".

As an example, the model with input that includes frequency and vector encoding is represented as E_{I_0,S_0,F_1,V_1}. All 16 different combinations can be seen in Table 1. All combinations are explored in Section 3.

It should be mentioned that later in our experiments, specific symbols that have been discarded from the text, "uh" and "um", will be added to the text to add more language features to the pipeline and study their effects.

2.3. Modeling

In this section, we present the modeling for our two different approaches. In the first case, a classifier based on cross-entropy is introduced. In the second case, a contrastive learning approach is utilized. Both models are formulated within a multi-task setup.

2.3.1. Model for In-Text Encoding Scheme

In this section, the model for training and inference using in-text encoding is described. In order to leverage the pre-trained LLMs, the BERT-base-uncased model [36] serves as our base model to extract features for the classification layers. These layers consist of a dropout layer and two linear layers. Additionally, a secondary network that processes frequency vectors is laid out. This network consists of linear layers in an auto-encoder format. This network serves as a regularizer for the main network. The model is depicted in Figure 1. The primary network makes use of a classification loss, and the secondary network incorporates a regularization loss. As shown in [37], several architectures were considered to obtain the best fit for this task.

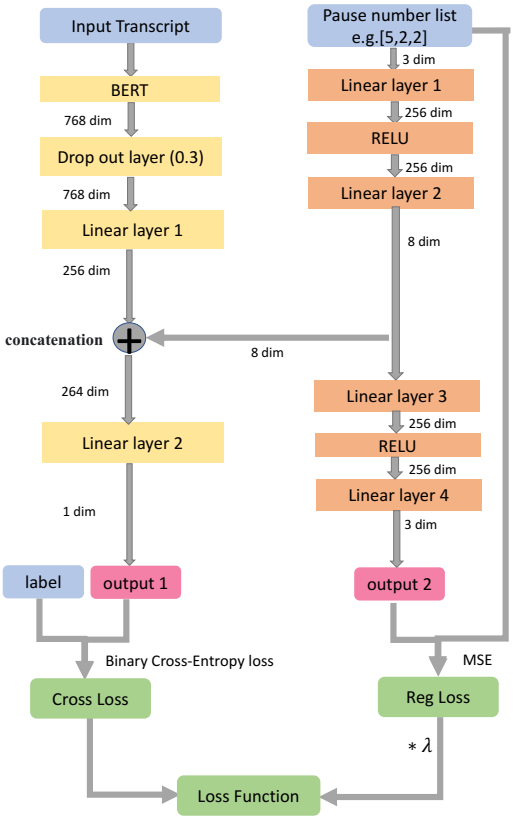

Figure 1. Model architecture for in-text encoding scheme.

For the classification loss, cross-entropy is used as follows:

$$\mathcal{L}_{Classification} = \sum_{i=1}^{N} CE(y_i, \hat{y}_i), \tag{1}$$

where $CE, y_i, \hat{y}_i$, and N are the cross-entropy loss, ground-truth label, predicted label, and number of samples, respectively. For the regularization loss, the mean square error (MSE) loss is used as follows:

$$\mathcal{L}_{Regularization} = MSE(V_F, \hat{V}_F), \tag{2}$$

where V_F and $\hat{V}_F$ are the true frequency input and its reconstruction, respectively. So, the total optimization cost becomes

$$\mathcal{L} = \mathcal{L}_{Classification} + \lambda \mathcal{L}_{Regularization}, \tag{3}$$

where λ is a hyperparameter to be optimized.

It should be noted that the secondary network is present for all the experiments, except for the baseline B_0.

2.3.2. Contrastive Learning

In this section, we propose using CL for classification. CL is a well-established approach that employs pairs of positive and negative examples to guide the deep learning model. Over time, CL has gained significant traction, surpassing earlier benchmarks and enhancing performance in fully supervised, semi-supervised, and self-supervised learning environments.

The core of CL is the creation of meaningful representations through the comparison of positive and negative instance pairings. The essential principle is that in an embedded space, similar examples should be close together, while distinct instances should be farther apart [30]. CL helps models find relevant features and similarities in the dataset by approaching learning as a differentiation process [30].

In this paper, we use an approach that the authors in [31] proposed. The authors proposed the dual contrastive learning (DualCL) framework that is a method that concurrently trains on the features of input samples and the parameters of classifiers. Essentially, DualCL treats the classifier parameters as extended samples linked to various labels. It then leverages CL to draw comparisons between these input samples and the extended, or augmented, samples. This approach allows for a more integrated and holistic learning process, where both the sample features and classifier parameters are understood and developed in relation to each other.

In this section, we address the modeling portion of the experiments. The notation from [31] is adopted. The focus is on a text classification task encompassing K different classes. The dataset under consideration, denoted as $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, comprises N individual training instances. Each instance consists of an input sentence $\boldsymbol{x}_i \in \mathbb{R}^L$, comprising L words, alongside its corresponding label y_i. For clarity, the study uses $\mathcal{I} = \{1, 2, \cdots, N\}$ to represent the index set of the training samples and $\mathcal{K} = \{1, 2, \cdots, K\}$ to denote the index set of the labels.

We explore self-supervised contrastive learning. This technique involves a dataset of N training samples, each accompanied by at least one augmented version within the set. If $j(i)$ represents the index of the augmented sample originating from the i-th sample, the standard formula for contrastive loss is as follows:

$$\mathcal{L}_{\text{self}} = \frac{1}{N} \sum_{i \in \mathcal{I}} -\log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_{j(i)} / \tau\right)}{\sum_{a \in \mathcal{A}_i} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \tag{4}$$

where $\boldsymbol{z}_i$ signifies the normalized form of $\boldsymbol{x}_i$. The set $\mathcal{A}_i := \mathcal{I} \backslash \{i\}$ is the contrastive samples' set. The dot product is represented by the symbol $\cdot$, and $\tau > 0$ acts as the temperature factor. In this context, the i-th sample is labeled as the anchor, the $j(i)$-th sample is considered a positive sample, and the rest of the samples, totaling $N - 2$, are deemed negative in relation to the i-th sample.

In the context of feature representation $\boldsymbol{z}_i$ and classifier $\boldsymbol{\theta}_i$ for a given input example x_i, the goal is to align the softmax transformation of $\boldsymbol{\theta}_i \cdot \boldsymbol{z}_i$ with the actual label of $\boldsymbol{x}_i$. The

column of $\boldsymbol{\theta}_i$ that corresponds to the true label of $\boldsymbol{x}_i$ is represented as $\boldsymbol{\theta}_i^*$. The objective is to maximize the dot product $\boldsymbol{\theta}_i^* \cdot \boldsymbol{z}_i$, thereby enhancing the representation of both $\boldsymbol{\theta}_i$ and $\boldsymbol{z}_i$ with supervised learning.

To achieve this, the dual contrastive loss is introduced. This loss function aims to maximize the dot product $\boldsymbol{\theta}_i^* \cdot \boldsymbol{z}_j$ when $\boldsymbol{x}_j$ shares the same label as $\boldsymbol{x}_i$, and minimize it when $\boldsymbol{x}_j$ has a different label. This approach is designed to leverage the relationships between various training samples, effectively distinguishing between those with similar and dissimilar labels.

For a given anchor $\boldsymbol{z}_i$, derived from the input $\boldsymbol{x}_i$, we categorize $\left\{\boldsymbol{\theta}_j^*\right\}_{j\in\mathcal{P}_i}$ as positive samples and $\left\{\boldsymbol{\theta}_j^*\right\}_{j\in\mathcal{A}_i\setminus\mathcal{P}_i}$ as negative samples. The contrastive loss is then defined as

$$\mathcal{L}_z = \frac{1}{N}\sum_{i\in\mathcal{I}}\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} -\log\frac{\exp\left(\boldsymbol{\theta}_p^*\cdot\boldsymbol{z}_i/\tau\right)}{\sum_{a\in\mathcal{A}_i}\exp(\boldsymbol{\theta}_a^*\cdot\boldsymbol{z}_i/\tau)}. \tag{5}$$

In this formula, τ is a positive real number that serves as the temperature factor. The set $\mathcal{A}_i := \mathcal{I}\setminus\{i\}$ includes the indices of all contrastive samples, while $\mathcal{P}_i := \{p \in \mathcal{A}_i : y_p = y_i\}$ denotes the set of positive sample indices. The term $|\mathcal{P}_i|$ represents the size or cardinality.

To continue, with $\boldsymbol{\theta}_i^*$ as the anchor, $\{\boldsymbol{z}_j\}_{j\in\mathcal{P}_i}$ are considered positive samples, and $\{\boldsymbol{z}_j\}_{j\in\mathcal{A}_i\setminus\mathcal{P}_i}$ are negative samples. This forms the basis for another type of contrastive loss, defined as

$$\mathcal{L}_\theta = \frac{1}{N}\sum_{i\in\mathcal{I}}\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} -\log\frac{\exp(\boldsymbol{\theta}_i^*\cdot\boldsymbol{z}_p/\tau)}{\sum_{a\in\mathcal{A}_i}\exp(\boldsymbol{\theta}_i^*\cdot\boldsymbol{z}_a/\tau)} \tag{6}$$

The dual contrastive loss then combines these two contrastive loss terms:

$$\mathcal{L}_{\text{Dual}} = \mathcal{L}_z + \mathcal{L}_\theta \tag{7}$$

In the joint training and prediction phase, the goal is to ensure that $\boldsymbol{\theta}_i$ is an effective classifier for $\boldsymbol{z}_i$. This is achieved using a modified cross-entropy loss, designed to maximize $\boldsymbol{\theta}_i^*\cdot\boldsymbol{z}_i$ for each input $\boldsymbol{x}_i$:

$$\mathcal{L}_{\text{CE}} = \frac{1}{N}\sum_{i\in\mathcal{I}} -\log\frac{\exp(\boldsymbol{\theta}_i^*\cdot\boldsymbol{z}_i)}{\sum_{k\in\mathcal{K}}\exp\left(\boldsymbol{\theta}_i^k\cdot\boldsymbol{z}_i\right)}, \tag{8}$$

where θ_i^k is the k-th column of $\boldsymbol{\theta}_i$.

To train the encoder f effectively, both training objectives are minimized simultaneously, enhancing the quality of the feature representations and the classifiers. The overall loss function is given as

$$\mathcal{L}_{\text{overall}} = (1-\alpha)\mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{Dual}}\,. \tag{9}$$

Here, α is a hyperparameter that modulates the impact of the dual contrastive loss in the overall training process. Later on in our experiments, $\mathcal{L}_{Regularization}$ will be added to the $\mathcal{L}_{\text{overall}}$ as follows:

$$\mathcal{L}_{\text{overall}} = (1-\alpha)\mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{Dual}} + \lambda\mathcal{L}_{Regularization}, \tag{10}$$

where λ is a hyperparameter to be optimized. This formulation will benefit from the secondary network shown in Figure 1. It represents a clear multi-task learning paradigm, from which the training can greatly benefit. The experiments in Section 3.3 demonstrate that this is indeed the case.

3. Results

In this section, we cover the impact of in-text encoding of pauses (Section 3.1) and filler words (Section 3.2). In Section 3.3, we study the effects of CL on the model performance. Our code for all the experiments can be found at https://github.com/ARoS-NCSU/Dementia-Detection-InTextEmbedding, accessed on 16 May 2024.

3.1. In-Text Pause Encoding Results

We study the effect of in-text encoding on model performance. The results are summarized in Table 2. The row for B_0 corresponds to the base model without any in-text encoding information available. All 16 combinations mentioned in Table 1 were explored. In all experiments, we used 20-fold cross-validation to evaluate the results. We used the Adam optimizer [38] with default parameters. For λ, we performed a grid search in the (0, 1] interval. After hyperparameter optimization, we chose 0.75. For each fold, all models were trained for 20 epochs and the best performance is reported.

As can be seen, for the E_{I_0,S_0,F_0,V_0} model, performance is around 0.58 and 0.56 for accuracy and f1-score, respectively. In this case, no in-text encoding is applied. For E_{I_1,S_0,F_0,V_0}, where in-text encoding is applied, we observe a drop in performance but still within the standard deviation of the model with no in-text encoding. This indicates that in-place encoding is not helpful on its own. For all other encodings, we observe a significant performance boost compared to the first two encodings. In particular, encoding E_{I_1,S_0,F_1,V_0}, where in-place pauses and their corresponding frequencies are combined, achieved the highest accuracy and f1-score, 0.84 and 0.85, respectively. In all experiments in Table 2, all artifacts and symbols are removed from the text, including "uh" and "um" filler words.

Table 2. In-text pause encoding results for 16 different combinations.

Input Type	Acc.	F1
E_{I_0,S_0,F_0,V_0}	0.58 ± 0.11	0.56 ± 0.24
E_{I_1,S_0,F_0,V_0}	0.50 ± 0.07	0.46 ± 0.25
E_{I_0,S_1,F_0,V_0}	0.83 ± 0.06	0.85 ± 0.05
E_{I_0,S_0,F_1,V_0}	0.83 ± 0.06	0.84 ± 0.06
E_{I_0,S_0,F_0,V_1}	0.81 ± 0.08	0.84 ± 0.06
E_{I_1,S_1,F_0,V_0}	0.83 ± 0.07	0.84 ± 0.07
E_{I_1,S_0,F_1,V_0}	**0.84 ± 0.06**	**0.85 ± 0.06**
E_{I_1,S_0,F_0,V_1}	0.82 ± 0.06	0.83 ± 0.06
E_{I_0,S_1,F_1,V_0}	0.83 ± 0.06	0.84 ± 0.05
E_{I_0,S_1,F_0,V_1}	0.82 ± 0.06	0.84 ± 0.05
E_{I_0,S_0,F_1,V_1}	0.82 ± 0.06	0.84 ± 0.06
E_{I_1,S_1,F_1,V_0}	0.83 ± 0.07	0.85 ± 0.06
E_{I_1,S_1,F_0,V_1}	0.83 ± 0.06	0.85 ± 0.05
E_{I_1,S_0,F_1,V_1}	0.83 ± 0.05	0.85 ± 0.04
E_{I_0,S_1,F_1,V_1}	0.82 ± 0.07	0.84 ± 0.06
E_{I_1,S_1,F_1,V_1}	0.83 ± 0.08	0.85 ± 0.07
$E_{Average}$	0.79	0.80
B_0	0.56 ± 0.11	0.42 ± 0.27

We also explored the effect of introducing single pauses (short, medium, or long), or a pair of them but did not observe a clear pattern that led us to believe that one combination was better than the others. Appendix A provides the details of this analysis.

3.2. Filler Word Encoding Results

We expand our previous analysis by considering the filler words "uh" and "um". It has been shown that filler word count plays a role in differentiating a dementia group from a control group [39–41]. The results are presented in three phases, as shown in Table 3.

Either the encoding for filler words is fixed while varying the encoding for pauses, or the encoding for pauses is fixed while varying the encoding for filler words. For phase 1, the encoding for the filler words to use $Filler(E_{I_0,S_0,F_0,V_0})$ is fixed, i.e., only the vector of counts is included, and the encoding for the pauses is varied. For phase 2, the encoding of the pauses to use $Pause(E_{I_1,S_1,F_1,V_1})$ is fixed, i.e., the pause in which all pause encoding is present, and the filler word encoding is varied. Finally, for phase 3, we fix the encoding of the filler words to use $Filler(E_{I_1,S_1,F_0,V_1})$ and vary the encoding for the pauses. The motivation for this analysis is to measure the impact of different choices of pause and filler word encoding.

Table 3. In-text filler word encoding results.

	Phase 1		**Phase 2**		**Phase 3**	
Input Type	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
E_{I_0,S_0,F_0,V_0}	0.53	0.41	0.83	0.85	0.84	0.85
E_{I_1,S_0,F_0,V_0}	0.57	0.46	0.83	0.85	0.84	0.86
E_{I_0,S_1,F_0,V_0}	0.82	0.85	0.84	0.86	0.83	0.85
E_{I_0,S_0,F_1,V_0}	0.82	0.84	0.83	0.85	0.83	0.86
E_{I_0,S_0,F_0,V_1}	0.84	0.86	0.84	0.85	0.82	0.84
E_{I_1,S_1,F_0,V_0}	0.84	0.86	0.84	0.85	0.84	0.86
E_{I_1,S_0,F_1,V_0}	0.84	0.86	0.84	0.85	0.83	0.84
E_{I_1,S_0,F_0,V_1}	0.84	0.86	0.84	0.85	0.83	0.85
E_{I_0,S_1,F_1,V_0}	0.83	0.85	0.81	0.84	0.82	0.84
E_{I_0,S_1,F_0,V_1}	0.83	0.85	0.83	0.85	0.83	0.85
E_{I_0,S_0,F_1,V_1}	0.82	0.84	0.83	0.85	0.83	0.85
E_{I_1,S_1,F_1,V_0}	0.82	0.84	0.82	0.84	0.84	0.85
E_{I_1,S_1,F_0,V_1}	0.83	0.85	0.84	0.86	0.82	0.84
E_{I_1,S_0,F_1,V_1}	0.82	0.84	0.82	0.84	0.85	0.86
E_{I_0,S_1,F_1,V_1}	0.84	0.85	0.81	0.84	0.82	0.84
E_{I_1,S_1,F_1,V_1}	0.83	0.86	0.84	0.85	0.83	0.84
$E_{Average}$	0.79	0.80	0.83	0.85	0.83	0.85

In phase 1, $Filler(E_{I_0,S_0,F_0,V_0})$ was used, where only the filler words (i.e., uh/um) were added to the text. Compared to the best results in Table 2, a 1% performance enhancement can be seen in terms of f1-score over multiple encodings. Although the average performance is the same as in Table 2, individual encodings showed better performance in most cases. This fact motivated the phase 2 experiments.

Given the improvements achieved by adding uh/um to the text, we started to encode them in the same way as for the pause information. In phase 2, a new I, S, F, V was introduced for the uh/um encoding. To find which of these combinations resulted in the best performance, first, a pause encoding, $Pause(E_{I_1,S_1,F_1,V_1})$, was fixed to perform this experiment. It can be observed that the $Filler(E_{I_1,S_1,F_0,V_1})$ and $Filler(E_{I_0,S_1,F_0,V_0})$ resulted in the best performance. Out of convenience, we chose $Filler(E_{I_1,S_1,F_0,V_1})$ moving forward. In phase 3, the filler word encoding was fixed to repeat the experiments for all 16 pause encoding experiments.

The main difference, in terms of performance, between phases 1 and 2 can be seen in encodings E_{I_0,S_0,F_0,V_0} and E_{I_1,S_0,F_0,V_0} (first two rows). There is roughly a 30% improvement across both metrics, which shows the effectiveness of the filler encoding. Also, due to these improvements, the average performance over all the encodings improved by 4% and 5% for accuracy and f1-score, respectively. The results in phase 3 are similar to phase 2, but $Pause(E_{I_1,S_0,F_1,V_1})$ achieved 85% and 86% in accuracy and f1-score, respectively, which is the best performance over all the phases.

In this section, we extended the work in [24] by exploring different aspects of the in-text pause encoding. Also, we incorporated filler words into text and found the optimal

setup for this approach, which resulted in significant performance improvement in some pause encodings and also improved model performance overall.

3.3. Contrastive Learning

In this section, the results of using CL to perform the classification are presented. For this, we used the DualCL framework as introduced in Section 2.3.2. For more clarity, samples in the same class are considered positive, while if they belong to different classes, they are considered negative samples. We used 20-fold cross-validation with 20 epochs for the evaluation of the model performance. We used the Adam optimizer [38] with default parameters. For hyperparameters λ and α, we chose 0.75 and 0.5 after performing a grid search in the range [0–1].

The first two rows in Table 4 show the impact of CL on the base model B_0. Compared to the baseline, B_0, the CL model achieved a roughly 30% improvement across both metrics, which shows the effectiveness of the CL approach. Keeping the filler words (third row) also improves accuracy, and f1-scores by an additional 1%. Compared to the previous results, we achieved a new best accuracy of 86%, which is an improvement of 1%.

Table 4. Results showing impact of contrastive learning (CL) and data augmentation (Aug) on different models.

Input Type	Acc.	F1
B_0	0.56 ± 0.11	0.42 ± 0.27
B_0 + CL	0.85 ± 0.04	0.84 ± 0.04
B_0 with Fillers + CL	0.86 ± 0.04	0.85 ± 0.04
B_0 + CL + Aug	0.87 ± 0.05	0.85 ± 0.05
B_0 with Fillers + CL + Aug	0.85 ± 0.05	0.84 ± 0.05
$Pause(E_{I_1,S_0,F_1,V_1}), Filler(E_{I_1,S_1,F_0,V_1})$	0.85 ± 0.06	0.86 ± 0.06
$Pause(E_{I_1,S_0,F_1,V_1}), Filler(E_{I_1,S_0,F_1,V_1})$ + CL	0.87 ± 0.08	0.86 ± 0.09

Since the dataset for training is small, we explored augmentation techniques to improve model performance. Similar augmentation techniques as used in previous sections were used, but they did not improve the model's performance. So, they are not reported in this paper. In Table 4 (rows 4 and 5), a contextualized embedding augmentation [42] was used to generate more samples. In particular, we used the BERT model for this contextualized augmentation. The choice of augmentation is very important, given the fact that too many changes can affect performance negatively due to the nature of the dataset and task. For training, three augmented samples per ground-truth sample were generated. In this case, when uh/um was not included, the model had a better performance in accuracy, better than all the results before. This might be due to the fact that augmentation imposes some changes, which in combination with uh/um impact the classification negatively. Compared to the results where no augmentation was utilized, the accuracy improved by 1% and the f1-score stayed the same.

Previously, we observed that combining pause and filler word encoding improved performance. It is only natural to combine these results with CL. For this, we chose the best result from Table 3 phase 3, namely, E_{I_1,S_0,F_1,V_1} ($Pause(E_{I_1,S_0,F_1,V_1})$ and $Filler(E_{I_1,S_1,F_0,V_1})$), to apply CL. Table 4 (last two rows) shows the results for this encoding. We observe a 2% increase in accuracy for this model.

In Table 5, we take a closer look at the impact of CL on pause encoding by repeating the experiments for Table 2 with CL present. An immediate conclusion that can be drawn from this table compared to Table 2 is that average accuracy has increased by 3%, which shows improvement in performance over all the encodings. In terms of the f1-score, the average stayed the same, which is due to some improvements but also some drops in performance. We also note that the variances in Table 2 are lower than the variance in Table 5. This is an indication that adding CL decreases the stability of the model.

Table 5. CL results for 16 combinations of pause encodings with no filler words.

Input Type	Acc.	F1
E_{I_0,S_0,F_0,V_0}	0.83 ± 0.08	0.82 ± 0.09
E_{I_1,S_0,F_0,V_0}	0.85 ± 0.08	0.84 ± 0.10
E_{I_0,S_1,F_0,V_0}	0.84 ± 0.08	0.81 ± 0.11
E_{I_0,S_0,F_1,V_0}	0.82 ± 0.08	0.80 ± 0.07
E_{I_0,S_0,F_0,V_1}	0.85 ± 0.08	0.83 ± 0.11
E_{I_1,S_1,F_0,V_0}	0.84 ± 0.09	0.82 ± 0.11
E_{I_1,S_0,F_1,V_0}	0.80 ± 0.10	0.78 ± 0.11
E_{I_1,S_0,F_0,V_1}	0.81 ± 0.09	0.79 ± 0.11
E_{I_0,S_1,F_1,V_0}	0.78 ± 0.08	0.79 ± 0.09
E_{I_0,S_1,F_0,V_1}	0.84 ± 0.09	0.82 ± 0.10
E_{I_0,S_0,F_1,V_1}	0.82 ± 0.08	0.80 ± 0.10
E_{I_1,S_1,F_1,V_0}	0.79 ± 0.08	0.77 ± 0.13
E_{I_1,S_1,F_0,V_1}	0.77 ± 0.09	0.72 ± 0.15
E_{I_1,S_0,F_1,V_1}	0.79 ± 0.09	0.77 ± 0.12
E_{I_0,S_1,F_1,V_1}	0.80 ± 0.09	0.79 ± 0.14
E_{I_1,S_1,F_1,V_1}	0.83 ± 0.07	0.81 ± 0.09
$E_{Average}$	0.82	0.80

3.4. Literature Comparison

In Table 6, we compared our results with some models in the literature that used the Pitt corpus dataset for classification. To the best of our knowledge, the results in [14] are state-of-the-art (SOTA) results. They thoroughly examined different choices for the input types (whole text or sentences), augmentation, pre-trained LLMs, and classifiers. After the investigations, they proposed using sentence-based rather than text-based pre-trained language models. The best model (S-BERT) uses sentences as input to the BERT-large model, with linear regression as the classifier. S-BERT achieved 0.88 and 0.87 in accuracy and f1-score, respectively. Our model is short 1% in performance in both metrics. In our modeling, we use the BERT-base model, which has far fewer parameters compared to BERT-large. Also, we used the whole text, not the sentences in each text, for classification. The authors in [14] also reported results for the text-based approach (T-BERT). Compared to their results, our model has a 2% improvement in both metrics. This shows our model is more efficient in processing whole text than the model proposed in [14]. In [15], the authors used a pre-trained LLM with RNNs. In [18,19], the authors used RNNs, attention mechanism, and Transformers to perform classification. It can be observed that models utilizing pre-trained LLMs achieve better results.

Table 6. Comparison of different methods using Pitt corpus dataset.

Input Type	Acc.	F1	No. Params.
S-BERT [14]	0.88	0.87	340M
T-BERT [14]	0.85	0.84	340M
Y. Pan et al. [19]	–	0.84	–
P. Saltz et al. [18]	0.76	0.76	110M
ALBERT+BiLSTM [15]	0.79	0.81	11M
Ours	0.87	0.86	110M

4. Discussion

In this work, we proposed using in-text encoding to improve the model's performance. We studied the effect of different types of pauses in Table A2 but did not see a significant difference in their performance. This may be due to not picking a relevant separation for the scale of pauses. Since filler words have proven to be an important indicator for dementia

detection, in Table 3, we investigated this fact and observed that for some pause encodings, the performance improved. It should be mentioned that the average performance over all the pause encoding was the same as the average in Table 2. In the next step, we introduced the filler word encoding in Table 3. The average accuracy and f1-score improved by 4% and 5%, respectively.

In the second portion of our modeling, we introduced DualCL to train the model. Table 4 shows the results for the cases where the filler words are absent or present. In the case where they are present, we can observe a 1% performance boost compared to the best prior results. Also, in Table 4, we used a contextualized augmentation to further improve the results. In the case where the filler words are not used, the model's accuracy increased by 1% compared to the best accuracy in Table 4 with no augmentation. It should be mentioned that augmentation was performed for all prior experiments, but the results were not satisfactory, and so, were not reported.

Lastly, we combined in-text encoding with the CL method. Table 5 shows that the average accuracy over all the pause encodings improved by 3%, which shows that this combination is very effective. It should be mentioned that the combination resulted in a higher standard deviation compared to other experiments. The last row in Table 3 shows the results for pause and filler encoding with the CL approach. This combination resulted in the best overall performance over both accuracy and f1-score.

The results in Table 6 show that our scheme is able to come close to SOTA performance (S-BERT [14]) with almost a third of the number of parameters. Since in our work, classification is performed at the document level, our result is directly comparable to T-BERT from [14]. This comparison shows an improvement over T-BERT, suggesting the effectiveness of our modeling scheme. However, further experiments are needed to confirm that the model appropriately attends to pause and filler word encodings. This is crucial for enhancing the model's interpretability, as these features are significant for clinicians.

5. Conclusions

In this paper, we proposed an in-text encoding methodology and the integration of contrastive learning (CL) as part of an LLM-based model for detection of dementia from speech transcripts. To the best of the authors' knowledge, the proposed approaches have not been explored by other authors in the literature. This work extends our previous research presented in [24], where we initially introduced in-text encoding. It is demonstrated that incorporating pauses and other language features (such as "uh" and "um" filler words) within the text model can considerably enhance performance. Additionally, it is shown that combining the CL approach with in-text encoding can further improve the model's performance. Overall, our modeling proved to be effective in distinguishing between the dementia and control groups.

The main limitation of our approach is related to automatic transcription. Accurate transcription with relatively precise timestamps is essential for the successful application of our method. For future work, we plan to incorporate other language features such as elongated words (e.g., "uhhh", "perrfect", etc.) within the text to study their effects, where the current language model may not properly encode them for analysis. For future work, incorporating an adaptive pause threshold within the models, exploring a combination of well-chosen augmentation techniques, and sentence-based pre-trained large language models could potentially improve model performance.

Author Contributions: Conceptualization, R.S.; methodology, R.S.; validation, R.S. and S.G.; formal analysis, R.S., S.G. and E.L.; investigation, R.S. and S.G.; resources, R.S. and S.G.; data curation, R.S. and S.G.; writing—original draft preparation, R.S., S.G. and E.L.; writing—review and editing, R.S., E.L., S.G., K.L.H. and A.J.; supervision, E.L., K.L.H. and A.J.; project administration, R.S., S.G., E.L., K.L.H. and A.J.; funding acquisition, E.L., K.L.H. and A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation (NSF) under awards IIS-1915599 and IIS-2037328.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are included in the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Effects of Pause Subsets

In Tables A1 and A2, the effect of each pause type is explored. To study how each type of pause affects the model performance, first, we removed each pause individually, and then, we just kept one pause at a time. To understand the effect of each pause on the model performance, we compared the best performance for each case to the average performance in Table 2. It should be mentioned that in the secondary network, the corresponding number associated with the removed pause frequency is set to zero.

If the short pauses are removed (the other two types are still present), the average accuracy and f1-score decrease by 1% and 2%, respectively. This shows that short pauses are important in dementia detection. In the case where the medium pauses are removed, the average accuracy is the same, but the average f1-score drops by 1%. Compared, to the short pause case, the effect is less severe. When the long pauses are removed, the results are similar to the medium case. To study the effect of each pause on performance individually, we address the case where only one pause is present in our analysis. First, if only short pauses are present, the model's performance drops by 1% and 2% for accuracy and f1-score, respectively. In the case of medium pauses, the model performance is preserved, which shows the significance of this type of pause. For long pauses, the accuracy is preserved, but there is a 1% decrease in f1-score.

Table A1. Effects of pause removal on average performance.

Pauses Included	Acc.	Δ Acc.	F1	Δ F1
All Pauses	0.79	-	0.80	-
Short Removed	0.78	−0.01	0.78	−0.02
Medium Removed	0.79	-	0.79	−0.01
Long Removed	0.79	-	0.79	−0.01
Short Only	0.78	−0.01	0.78	−0.02
Medium Only	0.79	-	0.80	-
Long Only	0.79	-	0.78	−0.02

Table A2. Studying the effect of each pause and their combinations on the model's performance.

Input Type	Short Removed		Medium Removed		Long Removed		Short Only		Medium Only		Long Only	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
E_{I_0,S_0,F_0,V_0}	0.53 ± 0.09	0.39 ± 0.30	0.53 ± 0.08	0.40 ± 0.28	0.53 ± 0.07	0.40 ± 0.30	0.51 ± 0.09	0.32 ± 0.35	0.58 ± 0.08	0.56 ± 0.26	0.52 ± 0.07	0.39 ± 0.33
E_{I_1,S_0,F_0,V_0}	0.50 ± 0.06	0.36 ± 0.31	0.53 ± 0.08	0.43 ± 0.20	0.55 ± 0.09	0.43 ± 0.27	0.51 ± 0.06	0.39 ± 0.30	0.52 ± 0.09	0.35 ± 0.34	0.52 ± 0.08	0.31 ± 0.31
E_{I_0,S_1,F_0,V_0}	0.84 ± 0.07	0.86 ± 0.06	0.83 ± 0.06	0.84 ± 0.06	0.83 ± 0.07	0.85 ± 0.06	0.82 ± 0.07	0.84 ± 0.07	0.83 ± 0.06	0.85 ± 0.05	0.83 ± 0.06	0.85 ± 0.06
E_{I_0,S_0,F_1,V_0}	0.82 ± 0.07	0.84 ± 0.06	**0.84 ± 0.06**	**0.86 ± 0.05**	0.81 ± 0.05	0.83 ± 0.05	0.83 ± 0.06	0.85 ± 0.05	0.82 ± 0.06	0.84 ± 0.05	0.83 ± 0.07	0.85 ± 0.06
E_{I_0,S_0,F_0,V_1}	0.83 ± 0.07	0.85 ± 0.06	0.83 ± 0.05	0.84 ± 0.05	0.83 ± 0.07	0.85 ± 0.06	0.81 ± 0.08	0.84 ± 0.06	0.84 ± 0.06	0.85 ± 0.06	0.82 ± 0.06	0.84 ± 0.05
E_{I_1,S_1,F_0,V_0}	0.81 ± 0.07	0.83 ± 0.06	0.83 ± 0.06	0.85 ± 0.06	0.83 ± 0.06	0.85 ± 0.05	0.84 ± 0.07	0.85 ± 0.07	0.83 ± 0.07	0.85 ± 0.06	0.81 ± 0.08	0.84 ± 0.07
E_{I_1,S_0,F_1,V_0}	0.81 ± 0.05	0.83 ± 0.04	0.84 ± 0.06	0.85 ± 0.06	0.81 ± 0.08	0.84 ± 0.06	0.82 ± 0.06	0.85 ± 0.06	0.82 ± 0.07	0.84 ± 0.07	0.83 ± 0.08	0.85 ± 0.07
E_{I_1,S_0,F_0,V_1}	0.81 ± 0.06	0.83 ± 0.05	0.82 ± 0.05	0.84 ± 0.04	0.83 ± 0.06	0.8 ± 0.05	0.83 ± 0.06	0.85 ± 0.06	0.82 ± 0.06	0.85 ± 0.05	0.82 ± 0.06	0.84 ± 0.06
E_{I_0,S_1,F_1,V_0}	0.81 ± 0.07	0.84 ± 0.06	0.83 ± 0.06	0.85 ± 0.05	0.83 ± 0.07	0.85 ± 0.06	0.82 ± 0.08	0.84 ± 0.07	0.81 ± 0.06	0.84 ± 0.04	0.82 ± 0.08	0.84 ± 0.07
E_{I_0,S_1,F_0,V_1}	**0.84 ± 0.06**	**0.85 ± 0.06**	0.82 ± 0.07	0.84 ± 0.06	0.82 ± 0.07	0.85 ± 0.06	0.82 ± 0.06	0.83 ± 0.05	0.83 ± 0.06	0.85 ± 0.06	0.83 ± 0.08	0.86 ± 0.07
E_{I_0,S_0,F_1,V_1}	0.82 ± 0.07	0.84 ± 0.06	0.83 ± 0.06	0.84 ± 0.06	0.83 ± 0.07	0.85 ± 0.06	0.81 ± 0.08	0.83 ± 0.07	0.83 ± 0.06	0.85 ± 0.05	0.82 ± 0.06	0.84 ± 0.05
E_{I_1,S_1,F_1,V_0}	0.82 ± 0.08	0.84 ± 0.07	0.83 ± 0.06	0.85 ± 0.05	0.83 ± 0.06	0.84 ± 0.06	0.82 ± 0.06	0.84 ± 0.05	0.82 ± 0.06	0.84 ± 0.05	0.83 ± 0.06	0.85 ± 0.06
E_{I_1,S_1,F_0,V_1}	0.83 ± 0.07	0.85 ± 0.06	0.81 ± 0.05	0.83 ± 0.05	0.83 ± 0.06	0.84 ± 0.06	0.82 ± 0.06	0.84 ± 0.05	0.82 ± 0.07	0.84 ± 0.06	0.82 ± 0.08	0.85 ± 0.06
E_{I_1,S_0,F_1,V_1}	0.83 ± 0.08	0.85 ± 0.06	0.83 ± 0.06	0.84 ± 0.05	0.81 ± 0.06	0.83 ± 0.05	0.81 ± 0.07	0.84 ± 0.05	0.82 ± 0.07	0.85 ± 0.06	0.83 ± 0.06	0.85 ± 0.05
E_{I_0,S_1,F_1,V_1}	0.82 ± 0.07	0.85 ± 0.06	0.82 ± 0.07	0.84 ± 0.06	0.83 ± 0.07	0.85 ± 0.07	0.83 ± 0.06	0.84 ± 0.05	0.84 ± 0.06	0.85 ± 0.06	0.82 ± 0.08	0.84 ± 0.06
E_{I_1,S_1,F_1,V_1}	0.82 ± 0.07	0.85 ± 0.06	0.82 ± 0.07	0.85 ± 0.06	0.81 ± 0.08	0.85 ± 0.07	0.83 ± 0.06	0.85 ± 0.05	0.84 ± 0.06	0.86 ± 0.06	0.83 ± 0.06	0.84 ± 0.06
$E_{Average}$	0.78	0.78	0.79	0.79	0.79	0.79	0.78	0.78	0.79	0.80	0.79	0.78

References

1. A.D. International. Dementia Statistics. Available online: https://www.alz.co.uk/research/statistics (accessed on 25 May 2024).
2. Yiannopoulou, K.G.; Papageorgiou, S.G. Current and future treatments in Alzheimer disease: An update. *J. Cent. Nerv. Syst. Dis.* **2020**, *12*, 1179573520907397. [CrossRef] [PubMed]
3. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
4. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
5. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
6. Chandra, R.; Kulkarni, V. Semantic and sentiment analysis of selected Bhagavad Gita translations using BERT-based language framework. *IEEE Access* **2022**, *10*, 21291–21315. [CrossRef]
7. Qu, C.; Yang, L.; Qiu, M.; Croft, W.B.; Zhang, Y.; Iyyer, M. BERT with history answer embedding for conversational question answering. In Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 1133–1136.
8. Hakala, K.; Pyysalo, S. Biomedical named entity recognition with multilingual BERT. In Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Hong Kong, China, 4 November 2019; pp. 56–61.
9. Thomas, C.; Keselj, V.; Cercone, N.; Rockwood, K.; Asp, E. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In Proceedings of the IEEE International Conference Mechatronics and Automation, Niagara Falls, ON, Canada, 29 July–1 August 2005; Volume 3, pp. 1569–1574. [CrossRef]
10. Radanovic, M.; Carthery-Goulart, M.T.; Charchat-Fichman, H.; Herrera, E., Jr.; Lima, E.E.P.; Smid, J.; Porto, C.S.; Nitrini, R. Analysis of brief language tests in the detection of cognitive decline and dementia. *Dement. Neuropsychol.* **2007**, *1*, 37–45. [CrossRef]
11. Murray, R.; Koenig, P.; Antani, S.; McCawley, G.; Grossman, M. Lexical acquisition in progressive aphasia and frontotemporal dementia. *Cogn. Neuropsychol.* **2007**, *24*, 48–69. [CrossRef]
12. Yuan, J.; Bian, Y.; Cai, X.; Huang, J.; Ye, Z.; Church, K. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; Volume 2020, pp. 2162–2166.
13. Valsaraj, A.; Madala, I.; Garg, N.; Baths, V. Alzheimer's dementia detection using acoustic & linguistic features and pre-trained BERT. In Proceedings of the 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI), Cairo, Egypt, 26–27 November 2021; pp. 171–175.
14. Roshanzamir, A.; Aghajan, H.; Soleymani Baghshah, M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 1–14. [CrossRef]
15. Nambiar, A.S.; Likhita, K.; Pujya, K.V.S.S.; Gupta, D.; Vekkot, S.; Lalitha, S. Comparative study of Deep Classifiers for Early Dementia Detection using Speech Transcripts. In Proceedings of the 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 24–26 November 2022; pp. 1–6. [CrossRef]
16. Cai, H.; Huang, X.; Liu, Z.; Liao, W.; Dai, H.; Wu, Z.; Zhu, D.; Ren, H.; Li, Q.; Liu, T.; et al. Multimodal Approaches for Alzheimer's Detection Using Patients' Speech and Transcript. In Proceedings of the International Conference on Brain Informatics, Hoboken, NJ, USA, 1–3 August 2023; Springer: New York, NY, USA, 2023; pp. 395–406.
17. Guo, Z.; Liu, Z.; Ling, Z.; Wang, S.; Jin, L.; Li, Y. 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, Barcelona, Spain, 8–13 December 2020; pp. 6161–6171.
18. Saltz, P.; Lin, S.Y.; Cheng, S.C.; Si, D. Dementia Detection using Transformer-Based Deep Learning and Natural Language Processing Models. In Proceedings of the 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), Victoria, BC, Canada, 9–12 August 2021; pp. 509–510. [CrossRef]
19. Pan, Y.; Mirheidari, B.; Reuber, M.; Venneri, A.; Blackburn, D.; Christensen, H. Automatic hierarchical attention neural network for detecting AD. In Proceedings of the Interspeech 2019. International Speech Communication Association (ISCA), Graz, Austria, 15–19 September 2019; pp. 4105–4109.
20. Afzal, S.; Maqsood, M.; Nazir, F.; Khan, U.; Aadil, F.; Awan, K.M.; Mehmood, I.; Song, O.Y. A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. *IEEE Access* **2019**, *7*, 115528–115539. [CrossRef]
21. Mirheidari, B.; Blackburn, D.; O'Malley, R.; Venneri, A.; Walker, T.; Reuber, M.; Christensen, H. Improving Cognitive Impairment Classification by Generative Neural Network-Based Feature Augmentation. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 2527–2531.
22. Jain, V.; Nankar, O.; Jerrish, D.J.; Gite, S.; Patil, S.; Kotecha, K. A novel AI-based system for detection and severity prediction of dementia using MRI. *IEEE Access* **2021**, *9*, 154324–154346. [CrossRef]
23. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A survey of data augmentation approaches for NLP. *arXiv* **2021**, arXiv:2105.03075.
24. Soleimani, R.; Guo, S.; Haley, K.; Jacks, A.; Lobaton, E. Dementia Detection by In-Text Pause Encoding. Preprint. 2024. Available online: https://www.preprints.org/manuscript/202408.0727/v1 (09-08-2024).

25. Luz, S.; Haider, F.; de la Fuente, S.; Fromm, D.; MacWhinney, B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020.
26. Becker, J.T.; Boiler, F.; Lopez, O.L.; Saxton, J.; McGonigle, K.L. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Arch. Neurol.* **1994**, *51*, 585–594. [CrossRef] [PubMed]
27. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the European Conference on Computer Vision, Springer: New York, NY, USA, 2020, pp. 474–490.
28. Harutyunyan, H.; Khachatrian, H.; Kale, D.C.; Ver Steeg, G.; Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **2019**, *6*, 96. [CrossRef] [PubMed]
29. Cramér, H.; Wold, H. Some theorems on distribution functions. *J. Lond. Math. Soc.* **1936**, *1*, 290–294. [CrossRef]
30. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **2020**, *9*, 2. [CrossRef]
31. Chen, Q.; Zhang, R.; Zheng, Y.; Mao, Y. Dual Contrastive Learning: Text Classification via Label-Aware Data Augmentation. *arXiv* **2022**, arXiv:2201.08702.
32. Vigo, I.; Coelho, L.; Reis, S. Speech-and language-based classification of Alzheimer's disease: A systematic review. *Bioengineering* **2022**, *9*, 27. [CrossRef]
33. He, R.; Chapin, K.; Al-Tamimi, J.; Bel, N.; Marquié, M.; Rosende-Roca, M.; Pytel, V.; Tartari, J.P.; Alegret, M.; Sanabria, A.; et al. Automated Classification of Cognitive Decline and Probable Alzheimer's Dementia Across Multiple Speech and Language Domains. *Am. J. Speech-Lang. Pathol.* **2023**, *32*, 2075–2086. _AJSLP-22-00403 [CrossRef]
34. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356.
35. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**, arXiv:2006.11477.
36. Hugging Face. Available online: https://huggingface.co/google-bert/bert-base-uncased (accessed on 16 May 2024).
37. Guo, S. Enhancing Dementia Detection in Text Data through NLP by Encoding Silent Pauses. 2024. Available online: https://repository.lib.ncsu.edu/items/7d00284d-35b0-4a9d-8b26-845157739f0f (accessed on 16 May 2024).
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
39. Karlekar, S.; Niu, T.; Bansal, M. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. *arXiv* **2018**, arXiv:1804.06440.
40. Wieling, M.; Grieve, J.; Bouma, G.; Fruehwald, J.; Coleman, J.; Liberman, M. Variation and change in the use of hesitation markers in Germanic languages. *Lang. Dyn. Chang.* **2016**, *6*, 199–234. [CrossRef]
41. Tottie, G. Uh and um as sociolinguistic markers in British English. *Int. J. Corpus Linguist.* **2011**, *16*, 173–197. [CrossRef]
42. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. *arXiv* **2018**, arXiv:1805.06201.