# GMR-7179

# Improving Alzheimer's Detection via Synthetic Data Generation Using GPT-4 and Multi-Level Embeddings

## Abstract

This study leverages large language models (LLMs), particularly GPT-4, to overcome the data limitations often encountered in Alzheimer's detection. We utilize GPT-4 for data augmentation, generating synthetic speech transcripts to enhance machine learning model training. Our approach combines fine-tuned BERT embeddings with CLAN-derived linguistic features, as well as sentence-level embeddings, to improve classification performance on the ADReSS2020 dataset. BERT and CLAN features capture detailed linguistic variants, while sentence embeddings offer robust semantic representations, collectively enhancing the accuracy and generalization of the models. Among the classifiers tested, the Random Forest model shows the best performance, achieving an accuracy of 88% with sentence embeddings, surpassing other models in detecting Alzheimer's from speech patterns. The integration of LLM-augmented data and multilevel embeddings presents a promising solution to the data scarcity issue in medical research, enabling more accurate and reliable Alzheimer's diagnoses.

## Introduction

Alzheimer's disease is a serious brain disorder affecting many individuals globally. While dementia is one of the main symptoms of Alzheimer's disease, it is important to recognize that many persons on the AD's continuum do not yet exhibit dementia but are in predetermined stages of the disease. Identifying the disease in early stage can mitigate its impact not only on dementia but also on other associated symptoms. There has been considerable research focused on the early detection of Alzheimer's disease, utilizing a range of machine learning techniques. However, the obstacle in training these machine learning models is due to lack of data. The insufficient data in the healthcare industry is due to data privacy concerns. To tackle this issue, the concept of data augmentation has been proposed. Data augmentation is a technique used to artificially increase the size and diversity of a dataset by creating modified versions of existing data. The objective of data augmentation is to apply transformation operations to the data to generate additional training samples, without transforming the original dataset. The augmented samples introduce variability into the dataset, which helps improve the robustness and generalization of machine learning models, particularly when working with small or sensitive datasets. This approach is especially valuable in medical research, where privacy concerns often restrict data sharing, allowing researchers to explore new insights without compromising confidentiality.

## Materials and Methods

**Dataset Overview**
We used the ADReSS2020 [1] challenge dataset for this study.

- It provides standardized, high-quality data of individuals describing the Cookie Theft picture, including both audio recordings and transcripts. In this experiment, we focused on transcripts.
- The dataset contains 1,955 speech segments from 78 non-AD subjects and 2122 speech segments from 78 AD subjects.

**Data Augmentation with LLM**
- We have implemented an LLM to generate more data using the original dataset.
- We have utilized OpenAI's GPT-4 model . By using the prompt, we have generated the synthetic data based on the original data (figure 1). Table I shows the prompt we used in this experiment.
- Our aim is to expand the dataset and compare the results.

Fig 1. Data augmentation using LLM

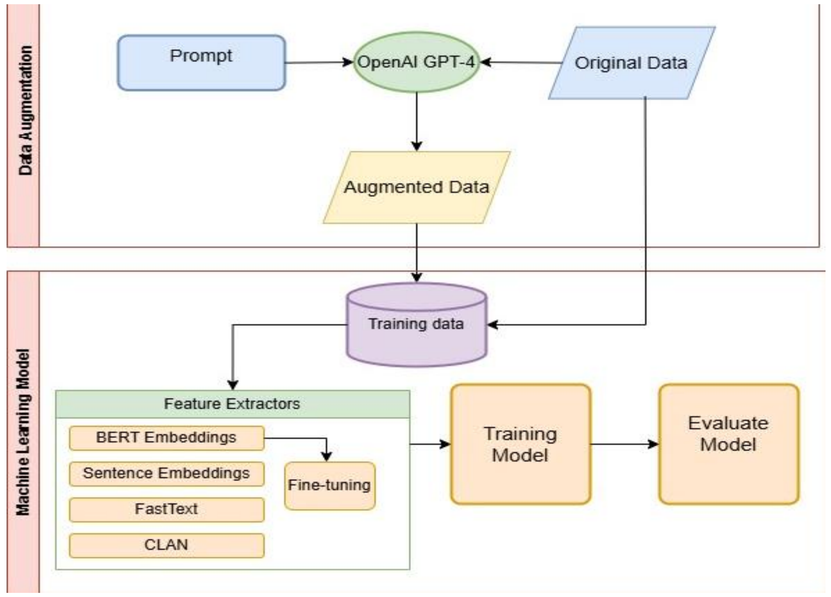


TABLE I. Prompt Used for Generating transcripts using LLM (GPT-4)

| |
|---|
| **Prompt:** |
| Generating transcripts based on the "Cookie Theft" picture description task. The goal is to expand the dataset for Alzheimer's Disease detection by creating transcripts that reflect natural speech patterns without introducing new errors. |
| • **Accuracy in Reproduction**: If the original transcript contains any grammar, punctuation errors, or disfluencies (such as pauses, repeated words, or hesitations), retain those errors in the new transcript. Do not introduce new errors or correct existing ones.<br>• **Minimal Modifications**: Make only very minor changes to the original transcript, such as slight rephrasing or using synonyms, ensuring the cognitive patterns and natural speech flow are preserved.<br>• **No New Errors**: Do not introduce any new grammar or punctuation mistakes unless they already exist in the original transcript. This will prevent the controlled patients' data from being altered, maintaining the accuracy of the dataset for the model. |
| **Original Transcript**: {original_transcript} |

**Implementation**
We describe two experiments conducted in this study.
- The first focuses on comparing different embeddings BERT, FastText, and BERT combined with CLAN features, with the baseline results of Haulcy et al [2] in the ADReSS dataset.
- The second experiment investigates the effect of using synthetic data generated by GPT-4, a large language model, for augmenting the dataset and evaluating its impact on classifier performance.

**Experiment 1: Comparison of BERT, FastText, and BERT+CLAN Embeddings**
We incorporated fine-tuned BERT embeddings and combined them with expanded CLAN-derived features, such as unique word count and average word length. Additionally, we used FastText embeddings to compare their effectiveness against BERT-based approaches. We have compared our results with the Baseline paper See Table 2.

**Experiment 2: Data Augmentation Using GPT-4**
We evaluated the performance of five classifiers on two dataset configurations, they are the dataset without data augmentation and an augmented dataset, which combines original data with synthetic data generated by GPT-4, a state-of-the-art Large Language Model developed by OpenAI.
The highest accuracy achieved in the baseline study was 85% using a Random Forest classifier. In contrast, we got a higher accuracy of 88% on the test set with data augmentation, as shown in Table 3.

## Results

**Comparison of BERT, FastText, and BERT+CLAN Embeddings:**
Overall, combining BERT embeddings with CLAN features improved performance in certain classifiers but did not lead to major increases inaccuracy compared to the baseline.

**Data Augmentation Using GPT-4:**
In particular, the augmented dataset yielded slight improvements in accuracy for the LDA and RF models when using sentence embeddings, demonstrating the effectiveness of data augmentation.

TABLE II. Test Set Accuracies For Classifiers, With An Additional Column For Our Results (Proposed).

| Features | Dim. Red. | Base LDA | Proposed LDA | Base DT | Proposed DT | Base 1NN | Proposed 1NN | Base SVM | Proposed SVM | Base RF | Proposed RF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BERT | None | 0.604 | **0.812** | 0.708 | **0.729** | 0.771 | 0.750 | 0.854 | 0.833 | 0.750 | **0.812** |
| | LDA(1) | 0.604 | **0.812** | 0.604 | **0.812** | 0.646 | **0.812** | 0.604 | **0.812** | 0.604 | **0.812** |
| | PCA(2) | 0.688 | 0.791 | 0.562 | **0.750** | 0.542 | **0.770** | 0.729 | 0.750 | 0.625 | **0.750** |
| | PCA(20) | 0.833 | 0.812 | 0.646 | **0.750** | 0.750 | **0.770** | 0.812 | 0.833 | 0.854 | 0.810 |
| BERT+CLAN | None | 0.729 | **0.812** | 0.750 | 0.729 | 0.771 | 0.541 | **0.812** | 0.583 | 0.812 | 0.729 |
| | LDA(1) | 0.729 | **0.812** | 0.708 | **0.812** | 0.708 | **0.812** | 0.708 | **0.812** | 0.708 | **0.812** |
| | PCA(20) | 0.729 | **0.812** | 0.708 | 0.729 | 0.667 | **0.770** | 0.771 | **0.833** | 0.792 | **0.812** |
| wordvectors | None | 0.813 | 0.645 | 0.688 | **0.708** | 0.667 | 0.666 | 0.500 | **0.687** | 0.833 | 0.791 |
| | LDA(1) | 0.813 | 0.645 | 0.750 | 0.687 | 0.771 | 0.687 | 0.813 | 0.666 | 0.750 | 0.687 |
| | PCA(2) | 0.729 | 0.729 | 0.542 | 0.666 | 0.500 | 0.645 | 0.500 | 0.708 | 0.667 | 0.708 |
| | PCA(70) | 0.812 | 0.708 | 0.562 | **0.729** | 0.688 | 0.666 | 0.500 | **0.791** | 0.771 | 0.750 |

TABLE III. Comparison of test set accuracies for Alzheimer's disease detection using BERT and Sentence Embeddings

| Algorithm | BERT Embeddings | | | Sentence Embeddings | |
|---|---|---|---|---|---|
| | Without Augumentation | LLM-Augmented Data | Baseline | Without Augumentation | LLM-Augmented Data |
| Decision Tree | 0.69% | 0.58% | 0.70% | 0.71% | 0.67% |
| LDA | 0.77% | 0.75% | 0.64% | 0.73% | **0.83%** |
| SVM (RBF Kernel) | 0.83% | **0.85%** | 0.85% | 0.81% | 0.77% |
| KNN(K=1) | 0.75% | 0.73% | 0.77% | 0.77% | 0.73% |
| KNN(K=5) | 0.79% | **0.83%** | - | 0.73% | **0.75%** |
| Random Forest | 0.75% | 0.75% | 0.75% | 0.79% | **0.88%** |

## Conclusions

This study demonstrated the effectiveness of using large language models (LLMs), particularly GPT-4, for augmenting datasets in Alzheimer's disease detection. Although we have already incorporated spontaneous speech datasets in this work, future research will focus on expanding the scope by integrating datasets with even greater variability in speech patterns and demographic representation. This will allow for more comprehensive testing of the model's ability to generalize across diverse populations and real-world scenarios. Additionally, we intend to explore the use of other large language models (LLMs) beyond GPT-4 to generate synthetic data. By comparing these models, we will investigate whether they lead to further improvements in classification accuracy and data diversity. This ongoing exploration of LLMs will enable us to address data scarcity challenges more effectively and push the boundaries of early Alzheimer's detection in clinical settings.

## Contact Information

Venkata Sai Bhargav Mutala vmutala@students.kennesaw.edu
Imaan Shahid ishahid1@students.Kennesaw.edu
Seyedamin Pouriyeh spouriye@kennesaw.edu

## References

[1]"Dementiabank.talkbank.org,"2020,accessed:2024-07-10.[Online].Available: https://dementia.talkbank.org/ADReSS-2020/

[2] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech, "Frontiers in Psychology, vol. 11, 2021. [Online]. Available: HTTPs://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.6241374

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author(s) : Venkata Sai Bhargav Mutala, Imaan Shahid
Advisors(s): Seyedamin Pouriyeh