ISSN 1870-4069

Node-weighted Graph Convolutional Network for Alzheimer's Dementia Detection from Transcribed Clinical Interviews with Data Augmentation

Lourdes Beatriz Cajica-Maceda, Perla Noemí Mendez-Zavaleta, Hugo Jair Escalante-Balderas, Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad

Instituto Nacional de Astrofísica, Óptica y Electrónica, Sta. María Tonantzintla, Puebla, México
{bcajica,perla.mendez,hugojair,ariel,fmartine}@inaoep.mx

Abstract. Early detection of Alzheimer's dementia from spontaneous speech is a critical task in clinical Natural Language Processing (NLP). However, existing datasets are often small and imbalanced, limiting the generalization of deep learning models. In this paper, we conduct an exploratory study on the effects of two data augmentation strategies, Synthetic Minority Oversampling Technique (SMOTE) and LLM-based generation, on an inductive Graph Convolutional Network (GCN) for dementia detection from transcribed clinical interviews. Experiments on the Pitt Corpus and ADReSS-2020 show that both techniques improve classification performance, with SMOTE consistently producing improvements and LLMs enhancing linguistic richness. Our findings support the use of augmentation approaches in low-resource neurocognitive diagnosis tasks.

Keywords: Convolutional Graph Network, Interviews-based Classification, Alzheimer's Dementia Detection.

1 Introduction

Dementia is a clinical syndrome characterized by a progressive decline in multiple cognitive functions, including memory, language, and activities of daily living, which ultimately affects independent life [9]. Despite significant research efforts, a definitive cure for disease-modifying therapies remains elusive, highlighting the importance of early and accurate prediction of Alzheimer's disease (AD) [11]. AD is the most common form of dementia, accounting for approximately 60–80% of cases, and is caused by various brain diseases and injuries, particularly neurodegenerative ones [9]. The clinical progression of dementia typically involves three stages: an initial presymptomatic phase, followed by mild cognitive impairment with short-term memory problems, and finally established dementia, which is marked by severe memory loss, disorientation, and sometimes neuropsychiatric symptoms [9].

Currently, the most reliable biomarkers for the diagnosis of dementia require advanced and often inaccessible resources. Neuroimaging techniques such as PET and MRI are expensive and not widely available, whereas cerebrospinal fluid tests are highly invasive [9, 11]. This reality has spurred the search for objective, non-invasive, and automated alternatives.

Analyzing a patient's speech and natural language has become a promising source of digital biomarkers with minimal clinical burden. Recent studies have explored the use of artificial intelligence to analyze transcripts and recordings of patients with dementia, identifying subtle linguistic markers of cognitive decline [1]. Therefore, applying Natural Language Processing (NLP) to patient interviews offers an accessible and cost-effective method for identifying early signs of dementia.

A significant challenge in this field is the scarcity of high-quality data, as the available corpora of patients with dementia are often small and imbalanced. To address this, standard techniques like Synthetic Minority Over-sampling Technique (SMOTE) are used to generate synthetic samples for balancing a dataset. More recently, the emergence of large language models (LLMs) has provided a new approach for generating realistic synthetic text, thereby further enriching training datasets [13]. These augmentation methods are crucial for mitigating data scarcity and improving the performance of classifiers.

In this context, graph-based architectures offer an efficient solution for low-resource settings. Graph Convolutional Networks (GCNs), in particular, can capture long-range semantic relationships at a low computational cost. Burdisso et al. [5] proposed a node-weighted inductive GCN model for depression detection in clinical interviews. This model improves upon standard GCNs by assigning learnable weights to each node's self-connection edges, relaxing the assumption that self-connections and neighbor edges are of equal importance. This node-weighted approach enhances interpretability while incorporating contextual information and has shown strong performance in low-resource classification tasks.

Motivated by these advantages, in this paper we adapt the node-weighted GCN architecture proposed by Burdisso et al. for the task of dementia detection from transcribed clinical interviews. In addition, we integrate and compare two data augmentation strategies: SMOTE and synthetic transcript generation using large language models (LLMs), with the aim of improving model robustness under data scarcity.

This paper is structured as follows. Section 2 reviews the relevant related work. Section 3 outlines our proposal, which encompasses a graph-based architecture, preprocessing strategies, and data augmentation techniques. Section 4 describes the datasets used in the experiments, specifically the Pitt Corpus and ADReSS-2020. Section 5 presents the experiments conducted using our proposal, while Section 6 discusses the obtained results. Finally, Section 7 provides our conclusions and some directions for future work.

2 Related Work

Several recent studies have explored the automatic detection of Alzheimer's disease using both traditional machine learning and deep learning techniques applied to linguistic features derived from interview transcripts. Qi et al. [14] provide a comprehensive review of noninvasive approaches for AD detection, highlighting the role of transcript-based linguistic markers in a wide range of machine learning and deep learning models. For instance, Balagopalan et al. [2] investigated transformer-based methods on the ADReSS dataset, achieving competitive results using only textual input. These works confirm the increasing interest in leveraging language-derived features from transcribed clinical interviews as a foundation for automated dementia classification.

The ADReSS 2020 shared task [12] provided a dataset extracted from the Pitt Corpus, aiming to classify speech samples from individuals diagnosed with AD and healthy controls. The top teams in this competition achieved F1-scores close to 0.89 using hand-crafted acoustic and linguistic characteristics [7]. Models such as RoBERTa, BERT with BiLSTM, and hybrid transformer-based networks have also been evaluated in subsets of the Pitt Corpus, reporting F1-scores between 79% and 95.5% depending on the architecture and pre-processing employed [7].

However, many of these models are dependent on large-scale training data to generalize effectively. Bouazizi et al. [4] discussed the limitations of applying large language models (LLMs) to dementia detection tasks, arguing that their high data requirements and potential biases make them less suitable for low-resource clinical settings. In contrast, they proposed generating synthetic training samples with GPT-3 to balance class distributions.

Although these methods demonstrate high performance, they often lack robustness when applied to new data distributions or when faced with highly imbalanced class scenarios [7, 4]. Furthermore, inconsistencies in dataset partitions, preprocessing pipelines, and evaluation criteria make it difficult to establish fair comparisons [7].

Graph-based approaches have emerged as a compelling alternative for text classification under low-resource conditions. Burdisso et al. [5] proposed a node-weighted inductive Graph Convolutional Network (GCN) for classifying clinical interviews, incorporating the structure of linguistic and document information through word-word and word-document relations, using Pointwise Mutual Information (PMI), TF-IDF, and PageRank-based weighting. In this work, the authors report results that consistently outperform those previously reported in the literature.

In the context of data augmentation, a recent study [8] explored several acoustic-based strategies for the detection of mild cognitive impairment (MCI) from spontaneous speech, comparing classical and generative methods. His findings reinforce the relevance of synthetic data generation, particularly with SMOTE and Generative Adversarial Networks (GANs), as an effective way to enhance classifier performance in low-resource medical domains.

From our literature review, we realized that to date, no previous study has systematically compared traditional oversampling techniques, such as SMOTE, with data generation using LLM within a graph-based framework for dementia detection, which motivates our proposal presented in the following section.

3 Proposed Methodology

As mentioned in the previous section, given the promising results of the node-weighted inductive Graph Convolutional Network (GCN) for classifying clinical interviews [5], this paper proposes adopting this approach for Alzheimer's Dementia Detection. We are looking to exploit three key strengths of the GCN approach: robustness in low-resource tasks, interpretability (since the weight allocation within the graph is transparent), and efficiency (as the graph is constructed only once). The preprocessing applied for the transcriptions is the same as that used in [5]. This includes converting all texts to lowercase, removing special characters and stopwords, and applying stemming using the *Snowball-Stemmer* algorithm.

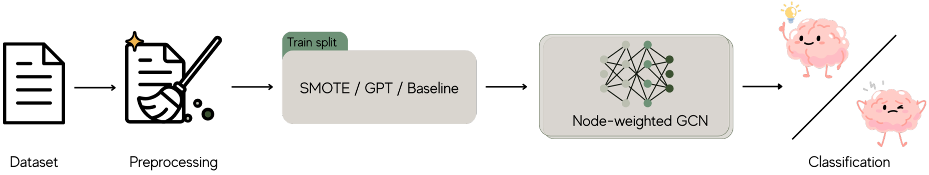

Fig. 1. Overview of our proposal.

Figure 1 describes the proposed methodology. We first preprocess the raw dataset and split it into training and test partitions. The training data, augmented with SMOTE, LLM, or without a strategy as a baseline, is used to construct a graph and train the node-weighted GCN. Finally, the trained GCN classifies the samples in the test partition.

3.1 Augmentation Strategies

For our methodology, we apply the following augmentation strategies:

SMOTE. The Synthetic Minority Oversampling Technique (SMOTE) [6] is a popular algorithm for balancing class distributions by generating synthetic samples of underrepresented classes. In the context of clinical interview transcripts, SMOTE works by taking feature representations of underrepresented interview recordings and interpolating between each minority example and its nearest neighbors in the feature space. The result is a set of synthetic interview feature vectors nearest to those of the existing samples. As a result, classifiers trained in SMOTE-augmented data learn more robust decision boundaries and exhibit better sensitivity to the minority class.

LLM-generated. We used GPT-4 and GPT-4.5 to generate synthetic interview transcripts [10]. The models were prompted with custom instructions and a few examples to simulate responses consistent with clinical interview scenarios, such as the Cookie Theft task (see Subsection 4.1). This method enables the creation of natural linguistic variations, allowing for semantically rich and context-aware text. The generated samples preserve the domain structure and linguistic realism, which can enhance the model's robustness. However, LLM-generated data may introduce artificial biases if the output is overly idealized or too consistent with training data patterns, potentially reducing variability.

4 Clinical Interview Datasets

This section describes the datasets used in our study for automatic dementia detection. We focus on transcribed clinical interviews provided by two well-known resources: the Pitt Corpus and the ADReSS-2020 dataset. In both cases, only the text of each intervention was used to build both the set of documents and the vocabulary.

4.1 Pitt Corpus

The Pitt Corpus [3] is part of the larger DementiaBank database within the TalkBank project. It comprises both audio recordings and transcriptions of clinical interviews. The corpus includes approximately 200 patients diagnosed with Alzheimer's disease and approximately 100 control individuals. It is structured around clinical tasks such as picture description (e.g., the "Cookie Theft" image). Although it is extensive, the corpus presents imbalances in class distribution and linguistic content, with differences in response lengths.

One of the most frequently used tasks in the Pitt Corpus is the description of the "Cookie Theft" picture, originally developed for the Boston Diagnostic Aphasia Examination (BDAE) by Harold Goodglass and Edith Kaplan in the 1970s. The image depicts a domestic scene: two children stand on a stool, taking cookies from a jar, while their mother is distracted by washing dishes as the sink overflows.

This task is widely employed in clinical linguistics and neuropsychology to elicit semi-structured spontaneous speech. It serves to evaluate multiple cognitive and linguistic abilities, including lexical access, syntactic complexity, narrative organization, and the ability to perceive and describe relevant visual elements.

In the context of dementia detection, the "Cookie Theft" task provides a controlled yet naturalistic setting where patients' language impairments become apparent. Individuals with Alzheimer's disease tend to produce shorter and less coherent descriptions, with more pauses, difficulty finding words, and omissions of key elements of the scene [7, 4].

4.2 ADReSS-2020

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset was introduced by [12] as a balanced subset of the Pitt Corpus. It includes a total of 156 participants: 78 with clinically confirmed Alzheimer's disease and 78 controls, matched by age and gender. Although class-balanced, the dataset contains systematic differences in response length that can affect models trained solely on transcripts [7].

Table 1. Experimental setups by dataset and augmentation method, with AD and non-AD sample counts.

Dataset	**Exp.**	**Strategy**	**AD**	**non-AD**
Pitt Corpus	1	Baseline	306	243
	2	SMOTE – balanced	306	306
	3	SMOTE – 200%	612	612
	4	LLM – balanced	306	306
	5	LLM – 200%	612	612
ADReSS-2020	6	Baseline	78	78
	7	SMOTE – 200%	156	156
	8	SMOTE – Pitt Corpus' size	306	306
	9	LLM – 200%	156	156
	10	LLM – Pitt Corpus' size	306	306

Table 2. Performance metrics for experiments 1-10 using 10-fold cross-validation.

Exp.	**Accuracy**	**F1-score (mean)**	**Precision**	**Recall**
1	0.74317 ± 0.1340	0.72084 ± 0.1712	0.78388 ± 0.1175	0.74317 ± 0.1340
2	$\mathbf{0.83103 \pm 0.0504}$	$\mathbf{0.83132 \pm 0.0507}$	$\mathbf{0.83875 \pm 0.0504}$	$\mathbf{0.83103 \pm 0.0504}$
3	0.83028 ± 0.0507	0.83078 ± 0.0508	0.83815 ± 0.0510	0.83028 ± 0.0507
4	0.82188 ± 0.0450	0.82178 ± 0.0466	0.83498 ± 0.0433	0.82188 ± 0.0450
5	0.79057 ± 0.0458	0.81676 ± 0.0476	0.81709 ± 0.0471	0.79057 ± 0.0458
6	0.90921 ± 0.0853	0.90685 ± 0.0874	0.92959 ± 0.0521	0.90921 ± 0.0853
7	0.91400 ± 0.0664	0.90892 ± 0.0759	0.92238 ± 0.0711	0.91400 ± 0.0664
8	0.91508 ± 0.0661	0.91084 ± 0.0738	0.92556 ± 0.0651	0.91508 ± 0.0661
9	$\mathbf{0.91783 \pm 0.0663}$	$\mathbf{0.91442 \pm 0.0724}$	$\mathbf{0.92791 \pm 0.0658}$	$\mathbf{0.91783 \pm 0.0663}$
10	0.90891 ± 0.0687	0.90199 ± 0.0816	0.91646 ± 0.0782	0.90891 ± 0.0687

Table 3. Performance metrics for experiments 6-10 using on the ADReSS-2020 standard train-test split.

Exp.	Accuracy	F1-score (mean)	Precision	Recall
6	0.87854 ± 0.0174	0.87819 ± 0.0175	0.88279 ± 0.0175	0.87854 ± 0.0174
7	0.87875 ± 0.0180	0.87843 ± 0.0181	0.88251 ± 0.0172	0.87875 ± 0.0180
8	0.87958 ± 0.0165	0.87916 ± 0.0167	0.88470 ± 0.0162	0.87958 ± 0.0165
9	**0.88083 ± 0.0188**	**0.88002 ± 0.0196**	**0.89014 ± 0.0150**	**0.88083 ± 0.0188**
10	0.86833 ± 0.0271	0.86777 ± 0.0276	0.87397 ± 0.0257	0.86833 ± 0.0271

5 Experiments

Our experiments aim to achieve two primary objectives. First, evaluating our methodology by applying an inductive GCN model, initially developed for detecting depression in clinical interview transcripts, to the task of detecting dementia using the same type of transcripts. Second, assess the impact of the two data augmentation techniques on the accuracy and robustness of our methodology in low-data scenarios. We ran ten experimental setups. For the Pitt Corpus, SMOTE and LLM-based augmentation were applied under two conditions: balanced sampling and 200% oversampling, and using the unmodified dataset as baseline. For the ADReSS-2020 dataset, which is already balanced, we evaluated 200% oversampling and scaling to match the Pitt Corpus size, and also using the unmodified dataset as baseline. Since the primary dataset does not provide a predefined train–test partition, we employ a 10-fold cross-validation for all training and evaluation. This procedure ensures robust and reliable performance estimates by exposing the model to every sample in both training and testing roles. For experiments on the ADReSS-2020 dataset, we also used the original train-test partition. Table 1 gives more details on the experimental design. For classification quality evaluation, we use accuracy, F1-score, precision and recall.

All experimental results are summarized in Tables 2 and 3. Table 2 presents the results using 10-fold cross-validation for Experiments 1-10, while Table 3 shows the results of experiments 6-10 on the standard ADReSS-2020 train–test split. The metrics are reported as mean ± standard deviation over 100 runs.

6 Discussion

The experimental results indicate that both SMOTE and LLM-based data augmentation strategies can improve model performance in the context of limited and imbalanced clinical interview data. On the Pitt Corpus, SMOTE yielded the most stable improvements, with accuracy increasing from 0.74317 to 0.83103 in the balanced configuration, as it is shown in the experiment 2 in Table 2. The use of LLMs also improved performance, although with slightly higher variance and marginally lower precision compared to SMOTE.

In the ADReSS-2020 dataset, using 10-fold cross validation, LLM-based augmentation achieved the highest accuracy of 0.91783 when expanding the dataset 200%. In this dataset, when using the original partition in training/testing sets, also LLM-based augmentation achieved the highest accuracy with 0.88083 when expanding the dataset 200%. This suggests that LLMs can provide realistic and semantically rich training samples that enhance generalization when balanced with the initial data. However, the improvements were modest compared to those on the Pitt Corpus, possibly due to the more controlled and balanced nature of the ADReSS-2020 set.

Overall, SMOTE proved more consistent, while LLM-based augmentation showed potential for generating semantically coherent and diverse samples, but may require better prompt design and post-generation filtering.

7 Conclusions

In this paper, through the proposed methodology, we demonstrated the versatility of the inductive GCN model, initially developed for depression detection, by adapting it to Alzheimer's disease detection. We also showed that data augmentation is useful for achieving reliable performance in low-resource settings.

On the Pitt Corpus, SMOTE improved accuracy, yielding stable and balanced enhancements. Meanwhile, the LLM-based augmentation obtained better accuracy for the ADReSS-2020 when scaled the dataset 200%.

Future work should explore hybrid augmentation pipelines that combine geometric (SMOTE) and generative methods (LLM), along with domain adaptation techniques to enhance robustness across diverse clinical datasets.

Acknowledgements. We acknowledge Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) and Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) for providing the necessary resources for the development of this research. We also thank the Pitt Corpus for delivering the data.

References

1. Al-Hammadi, M., Fleyeh, H., Åberg, A.C., Halvorsen, K., Thomas, I.: Machine learning approaches for dementia detection through speech and gait analysis: A systematic literature review. Journal of Alzheimer's Disease **100**(1), 1–27 (2024)
2. Balagopalan, A., Eyre, B., Rudzicz, F., Novikova, J.: To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. In: Interspeech 2020. pp. 2167–2171 (2020). https://doi.org/10.21437/Interspeech.2020-2557
3. 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. Archives of neurology **51**(6), 585–594 (1994)
4. Bouazizi, M., Zheng, C., Yang, S., Ohtsuki, T.: Dementia detection from speech: What if language models are not the answer? Information **15**(1), 2 (2023)

5. Burdisso, S., Villatoro-Tello, E., Madikeri, S., Motlicek, P.: Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In: Proceedings of Interspeech 2023. pp. 3617–3621 (2023)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**(1), 321–357 (2002). https://doi.org/10.1613/jair.953
7. Ding, K., Chetty, M., Noori Hoshyar, A., Bhattacharya, T., Klein, B.: Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. Artificial Intelligence Review **57**(12), 325 (2024)
8. Galban-Pineda, M.G.: Aumento de datos para detección de deterioro cognitivo en habla espontánea. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) (2024), unpublished manuscript (in Spanish)
9. Gunes, S., Aizawa, Y., Sugashi, T., Sugimoto, M., Rodrigues, P.P.: Biomarkers for alzheimer's disease in the current state: A narrative review. International Journal of Molecular Sciences **23**(9), 4962 (2022)
10. Holderried, F., Stegemann–Philipps, C., Herschbach, L., Moldt, J.A., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., Mahling, M.: A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. JMIR Medical Education **10**, e53961 (2024). https://doi.org/10.2196/53961
11. Huang, L., Yang, H., Che, Y., Yang, J.: Automatic speech analysis for detecting cognitive decline of older adults. Frontiers in Public Health **12**, 1417966 (2024)
12. Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., MacWhinney, B.: Alzheimer's dementia recognition through spontaneous speech (2021)
13. Nazir, A., Wang, Z.: A comprehensive survey of chatgpt: Advancements, applications, prospects, and challenges. Meta Radiology **1**(2), 100022 (2023)
14. Qi, X., Zhou, Q., Dong, J., Bao, W.: Noninvasive automatic detection of alzheimer's disease from spontaneous speech: a review. Frontiers in Aging Neuroscience **15**, 1224723 (2023). https://doi.org/10.3389/fnagi.2023.1224723