technologies

MDPI

Article

A Novel Methodology for Data Augmentation in Cognitive Impairment Subjects Using Semantic and Pragmatic Features Through Large Language Models

Luis Roberto García-Noguez [1], Sebastián Salazar-Colores [2], Siddhartha Mondragón-Rodríguez [3] and Saúl Tovar-Arriaga [1,*]

[1] Facultad de Ingeniería, Universidad Autónoma de Querétaro, Queretaro C.P. 76010, Mexico; luis.roberto.garcia@uaq.mx
[2] Centro de Investigaciones en Óptica, Leon C.P. 37150, Mexico; sebastian.salazar@cio.mx
[3] Instituto de Neurobiología, Universidad Nacional Autónoma de México Campus Juriquilla, Queretaro C.P. 76230, Mexico; sidmonrod@gmail.com
* Correspondence: saul.tovar@uaq.mx

Abstract

In recent years, researchers have become increasingly interested in identifying traits of cognitive impairment using audio from neuropsychological tests. Unfortunately, there is no universally accepted terminology system that can be used to describe language impairment, and considerable variability exists between clinicians, making detection particularly challenging. Furthermore, databases commonly used by the scientific community present sparse or unbalanced data, which hinders the optimal performance of machine learning models. Therefore, this study aims to test a new methodology for augmenting text data from neuropsychological tests in the Pitt Corpus database to increase classification and interpretability results. The proposed method involves augmenting text data with symptoms commonly present in subjects with cognitive impairment. This innovative approach has enabled us to differentiate between two groups in the database better than widely used text augmentation techniques. The proposed method yielded an increase in the metrics, achieving 0.8742 accuracy, 0.8744 F1-score, 0.8736 precision, and 0.8781 recall. It is shown that implementing large language models with commonly observed symptoms in the language of patients with cognitive impairment in text augmentation can improve the results in low-resource scenarios.

Keywords: cognitive impairment; low-resource scenarios; text augmentation; Pitt Corpus; memory impairment

Academic Editor: Lun Hu

Received: 27 June 2025
Revised: 29 July 2025
Accepted: 5 August 2025
Published: 7 August 2025

Citation: Garcia-Noguez, L.R.; Salazar-Colores, S.; Mondragón-Rodríguez, S.; Tovar-Arriaga, S. A Novel Methodology for Data Augmentation in Cognitive Impairment Subjects Using Semantic and Pragmatic Features Through Large Language Models. *Technologies* **2025**, *13*, 344. https://doi.org/10.3390/technologies13080344

1. Introduction

According to the World Health Organization (WHO), approximately 55 million people worldwide are living with dementia, and this number is projected to rise to 152 million by 2050. This syndrome can significantly affect the quality of life for those who are affected, as well as their families and caregivers. Individuals with dementia may experience changes in thinking, learning, memory, judgment, and decision making. Therefore, early detection of this syndrome is crucial. The initial stages of dementia often manifest as mild cognitive impairment (MCI). In specialized settings, the annual progression rate from MCI to dementia is about 10% [1].

Language impairment is a core feature of Alzheimer's disease (AD) and other dementias [2]. Prior studies have shown a link between AD symptom severity and declining speech and language capability in picture description tasks [3]. Speech and language changes include loss of written comprehension, naming disorders, fluent but empty speech, and semantic paraphasia [4]. Despite these changes, the techniques used for text augmentation in the classification of patients with dementia or mild cognitive impairment lack the careful application of these symptoms.

Given current limitations, natural language processing (NLP) is emerging as a novel method of assessing speech and language in individuals with neurologic and mental disorders. For example, in schizophrenia, the use of NLP techniques, such as latent semantic analysis, can identify features like incoherence [5]. Due to the promising results demonstrated in various natural language processing, proposals to analyze language using large language models (LLMs) in subjects with cognitive impairment have emerged. For example, Rosas et al. [6] developed a syntactic analysis that included filler words, pauses, formulated words, restarts, repetitions, fuzzy speech, and incomplete utterances. Using two classifiers, they achieve an accuracy of 86.42%. Moreover, Santander-Cruz et al. [7] classifies cognitive and non-cognitive subjects using semantic, demographic, lexical, and syntactic features. Using different machine learning models, they obtained an F1-score of 80%. In addition, Ramos-Morales et al. [8] proposed a model to analyze audio and text features reaching an F1-score of 86.42%. Likewise, the current language models that perform best require a large amount of data and are difficult to interpret [9]. Furthermore, obtaining a large amount of labeled data from patients with and without cognitive impairment is challenging due to the high cost and time required to collect the data.

Although some reported methods can increase the amount of text fed to machine learning models [10], applying these methods to neuropsychological tests may not generalize effectively or may introduce medical biases. This work proposes a new methodology for augmenting text data from neuropsychological tests in the Pitt Corpus database [11] to enhance classification results and achieve a more interpretable model. It is based on applying semantic and pragmatic features to augment text data using symptoms commonly present in subjects with cognitive impairment.

The essential contribution is an innovative methodology that employs large language models (LLMs) to create synthetic textual data guided by four salient linguistic features observed in individuals with cognitive impairment. This approach offers substantial advantages over existing textual data augmentation methods, yielding superior classification outcomes and greater interpretability for neuropsychological tests in the elderly population.

2. Related Work

Recently, studies have employed natural language processing (NLP) techniques in healthcare. For example, it has enabled the automatic extraction of information from unstructured data, including text, audio, and video, in medical records [12].

Because cognitive impairment can take several years to evolve from MCI to dementia, detection of cognitive impairment in the early stages is essential. Automatic detection emerges as a suitable strategy for this purpose because it eliminates the need for repeated consultations with a specialist. However, one of the main limitations for identifying traits associated with cognitive impairment using machine learning tools is that the databases generated for this purpose have limited data. New works have emerged to address this problem; for example, Masrani et al. [13] created a dataset comprising several thousand blog posts, including some from individuals with dementia and others from control subjects. They used this dataset to design a classifier based on Random Forests, k-nearest neighbors (KNNs), and artificial neural networks (ANNs), achieving an accuracy of 84%.

Some works have employed transfer learning using pre-trained word embeddings, given the specific data cases. For example, Chen et al. [14] applied Glove as a word embedding and mechanism attention, demonstrating the potency of automated speech analysis through transfer learning as a valuable tool for early screening of AD.

Recently, Lin and Washington [15] augmented text and audio data from the Pitt Corpus database. Their results indicate that synonym-based text data augmentation generally enhances the performance of the classification model. However, applying this method can generate false positives or negatives since there is no control over the changed words when creating samples.

Research classifying cognitive impairment subjects using NLP techniques has employed traditional data augmentation methods. However, these methods must consider the characteristics observed in patients with cognitive impairment to obtain better results and understand the features analyzed in the subject's samples.

3. Materials and Methods

This section presents the proposed methodology for text augmentation and the classification of subjects with cognitive impairment. It describes the dataset, text pre-processing, text data augmentation method, embedding model, and classification technique.

3.1. Dataset

As we said before, the Pitt Corpus database [11] was deployed in this study, which contains the Mini-Mental State Examination (MMSE) test, demographic data, audio recordings, and transcripts collected as part of the protocol administered by the Alzheimer's and Related Dementias' study at the University of Pittsburgh School of Medicine. The total participants comprised 243 healthy controls (HC), 307 people with mild cognitive impairment (MCI), possible Alzheimer's disease (AD), and other dementia diagnoses (Table 1). To reduce class imbalance and to analyze only one type of dementia, only subjects labeled as having possible and probable Alzheimer's disease were selected.

Table 1. Cognitive impairment and HC group comparisons of demographics and MMSE mean scores.

Group	Subjects	Gender	Age	MMSE	Education Years
HC	243	90M/153F	46–81	29	8–21
Cognitive impairment	307	118M/189F	49–90	20	6–20

The dataset contains the audio and manual transcripts from the cookie theft, fluency, and recall tasks. The present study focuses on cookie theft because it provides a standardized test used in various studies [16]. In this test, the subjects described everything they saw in the cookie theft picture [17]. This description is based on a familiar domestic scene, which requires the use of basic essential vocabulary learned in childhood.

The aim of applying this test evaluation is to gather information about patients' communication and attention deficits. We developed a method for augmenting text data in this dataset to increase the classification metrics between the two groups. The steps outlined in the following points were taken to achieve this.

3.2. Text Pre-Processing

As mentioned above, the database provides a set of speech samples classified into control and cognitive impairment subjects. Each speech recording was manually transcribed at the word level, following the TalkBank CHAT protocol [18]. These transcripts contain brief information about the test subject, including their identifier, age, gender, score achieved in the Mini-Mental State Examination (MMSE) test, and diagnosis. The

researcher's instructions and interventions are transcribed, as is the description provided during the test.

The corpus extracted from the transcript was generated by filtering, which only extracts the participant's response. The pre-processing applied to the corpus involves removing the CHAT symbology added by the researchers, punctuation marks, and special characters from the text, and converting it to lowercase (Figure 1).

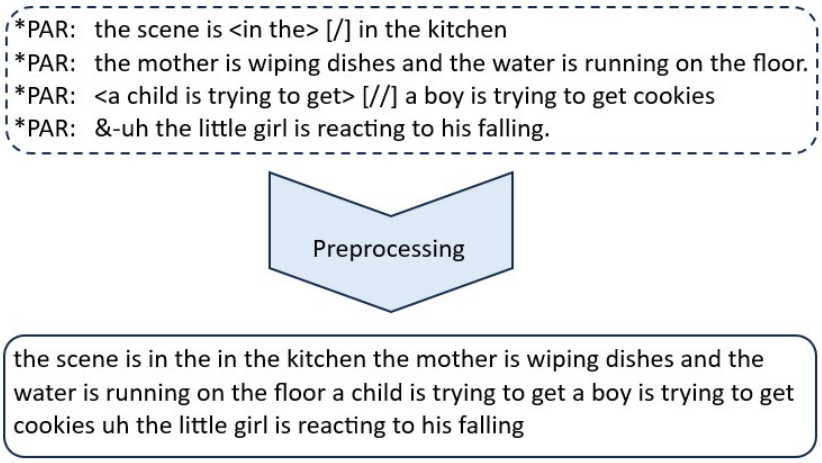

Figure 1. The samples underwent text pre-processing. At this stage, CHAT symbology, punctuation marks, and special characters were removed.

3.3. Text Augmentation

The proposed method for data augmentation involved three stages. The first requires analyzing the transcripts to identify whether they contain any of the four characteristics considered: incomplete ideas, circumlocutions, repetitious ideas, and illogical wording. These characteristics were considered because of the semantic and pragmatic elements present in the language of subjects with Alzheimer's disease in the initial stages [19]. The second and third stages involve selecting samples and performing data augmentation based on the features identified in the first stage. Figure 2 shows the steps of this process. These stages are described below.

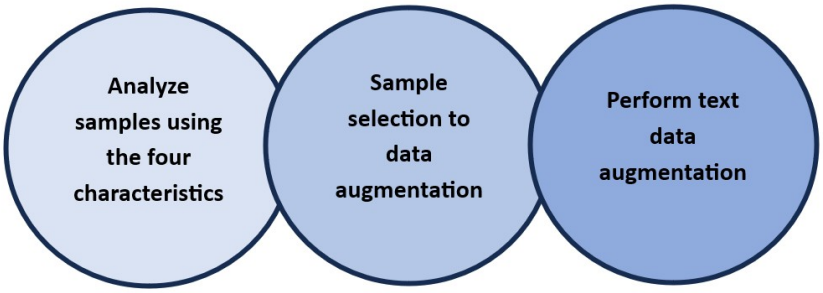

Figure 2. The text augmentation process. The first stage involves analyzing the samples, and the second and third stages are sample selection and data augmentation.

3.3.1. Analyzing Samples

To determine the characteristics observed in the language of both groups, the DeepSeek R1 7B model [20] was used to generate a description of each sample. This model was selected due to comparable results with larger models on different medical datasets [21].

DeepSeek R1 is a large language model developed by a research team in China. It performs well on complex mathematical, coding, and scientific reasoning tasks. The model is innovative, particularly in using reinforcement learning and model distillation. This

process utilizes chain-of-thought prompting, encouraging the model to "think out loud" or provide step-by-step reasoning in its responses. For example, when solving math problems, each step of the process will be shown. This approach enables the identification of mistakes and allows the model to self-evaluate and improve its accuracy through re-prompting or re-evaluating its steps. Using this model, samples were evaluated based on the four features proposed.

Because the image to be described by the subject presents a set of main ideas [22], the chain-of-thought approach helps to analyze the text step by step, considering the four proposed characteristics.

The samples were analyzed using one-shot prompting. A sample was proposed for each of the four characteristics analyzed (Figure 3).

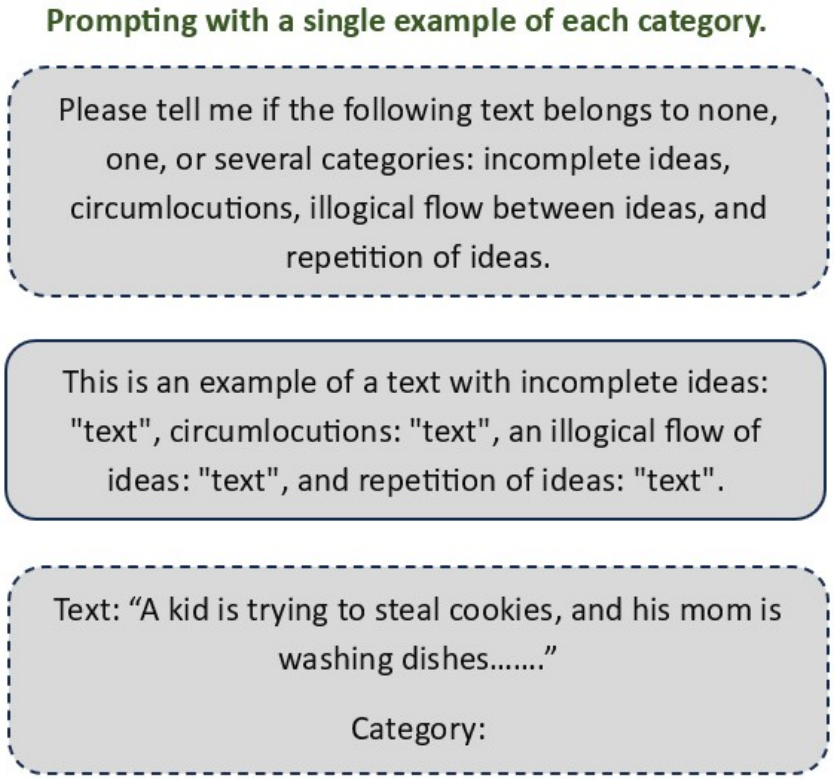

Figure 3. One shot prompt. A sample of each category was provided.

3.3.2. Sample Selection and Data Augmentation

Once each text was analyzed based on the proposed characteristics, new samples were generated, modifying the elements observed in the original text. In this way, the control samples were rephrased, but considering the characteristic of repeating ideas, the cognitive impairment samples were rephrased using incomplete ideas. These characteristics were selected because incomplete ideas were the most predominant characteristic in the samples of subjects with cognitive impairment, while the repetition of ideas was the least frequent. For both cases, the Llama 3.2 3B model [23] was used for rephrasing. This model was selected due to the promising results shown in text generation in medical applications [24]. This model was not fine-tuned due to the limited amount of available data.

Moreover, since not all samples contributed to improving the classifier's results, in the case of the control group samples, only the samples that did not have the characteristics of incomplete ideas or any other characteristics were augmented. In contrast, in the case of samples of cognitive impairment subjects, only the texts that did not have the characteristics of idea repetition or did not have any characteristics were augmented (Figure 4).

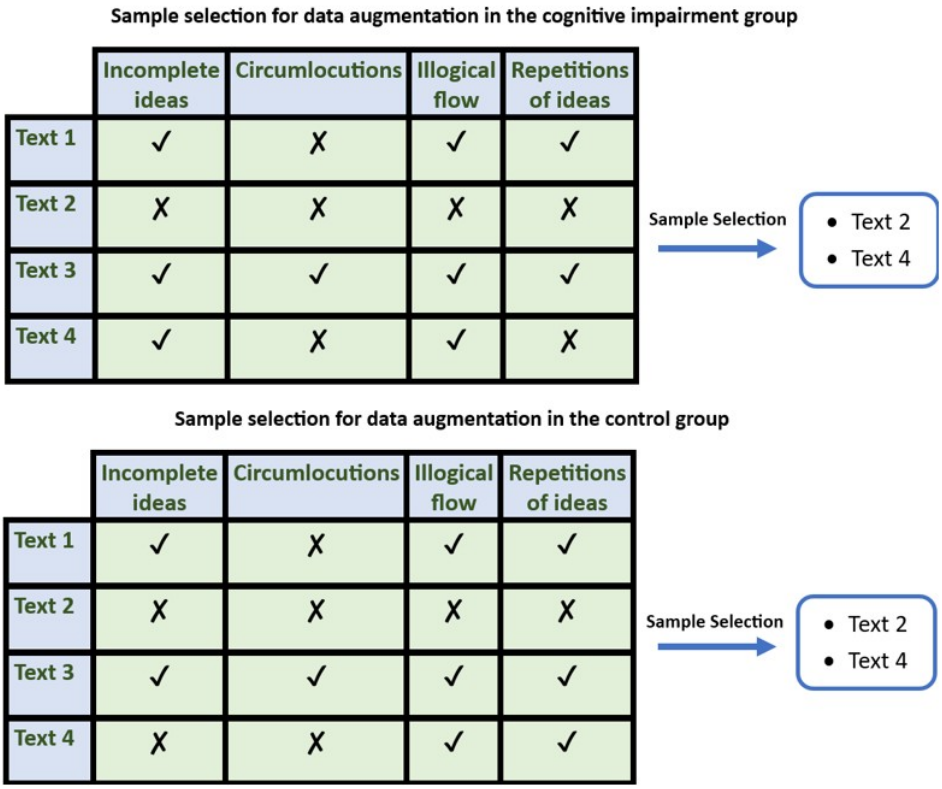

Figure 4. Samples from the group of subjects with cognitive impairment that do not have the characteristic of idea repetition and samples from the control group that do not have incomplete ideas will be selected to perform data augmentation.

3.4. Text Vectorization

The embedding models used to test the proposed method were BERT [25], RoBERTa [26], Linq-Embed-Mistral [27], and gte-Qwen2-7B-instruct [28]. Both BERT and RoBERTa were selected due to their architecture and performance in classification tasks [29]. Meanwhile, Linq-Embed-Mistral and gte-Qwen2-7B-instruct were selected because they were the two open-source models with the best performance in the MTEB benchmark [30] (21 May 2025). A detailed description of these embeddings is provided in the subsequent sections.

3.4.1. BERT and RoBERTa Model

The RoBERTa model [26] was selected to vectorize the samples. This method is based on the BERT model [25] but incorporates some improvements. On the one hand, BERT was pre-trained on two tasks. The first is masked language modeling (MLM), which predicts missing (masked) words in a sentence. The second is next-sentence prediction (NSP), where the system is trained to predict whether one sentence follows another. On the other hand, the RoBERTa model employs the same concept as BERT but eliminates the NSP component and utilizes larger batch sizes. Additionally, this method introduces dynamic masking for the MLM, whose masked tokens change during training epochs. The RoBERTa model maps words to a 768-dimensional space, just like BERT. A sentence vector was generated by averaging each word embedding. These vectors can be used for tasks such as clustering or semantic searches. In this case, they were used for the classification task.

3.4.2. Linq-Embed-Mistral Model

The Linq-Embed-Mistral model was built based on E5-mistral-7b-instruct and Mistral-7B-v0.1 [27]. This LLM incorporates key advances to improve text retrieval through advanced data refinement. This includes sophisticated data crafting, filtering, and teacher

model-guided negative mining to create high-quality triplet datasets. Linq-Embed-Mistral was selected for testing because it is one of the best models on the MTEB leaderboard [30]. This model generates an embedding of 4096 elements.

3.4.3. gte-Qwen2-7B-Instruct

The gte-Qwen2-7B-instruct model is the newest in the gte family of text embedding models, and it was selected because it currently holds the top ranking on the MTEB benchmark [30] for both English and Chinese. Built upon the Qwen2-7B large language model [28], it maintains the same training data and strategies as its predecessor, gte-Qwen1.5-7B-instruct. Key advancements include bidirectional attention, query-side instruction tuning, and extensive multilingual training on diverse data, enhancing its contextual understanding and applicability across various languages and tasks. The embedding has a dimension of 3584 elements.

3.5. Classifiers

We have applied a support vector machine and a logistic regression model as classification methods in this work. The choice of the methods is based on their results in the state-of-the-art techniques for similar tasks using word embeddings [31] and linguistic features [32]. These models are briefly described below.

3.5.1. Support Vector Machine

A support vector machine (SVM) is a supervised machine learning algorithm primarily used to classify tasks. It can also be used for regression and outlier detection. The core idea behind an SVM is to find the best possible hyperplane that separates data points into different classes. In 2D, this hyperplane is a line; in higher dimensions, it is a plane or a more complex surface. The optimal hyperplane is the one that maximizes the margin between the closest data points from different classes. These points are referred to as support vectors (Figure 5).

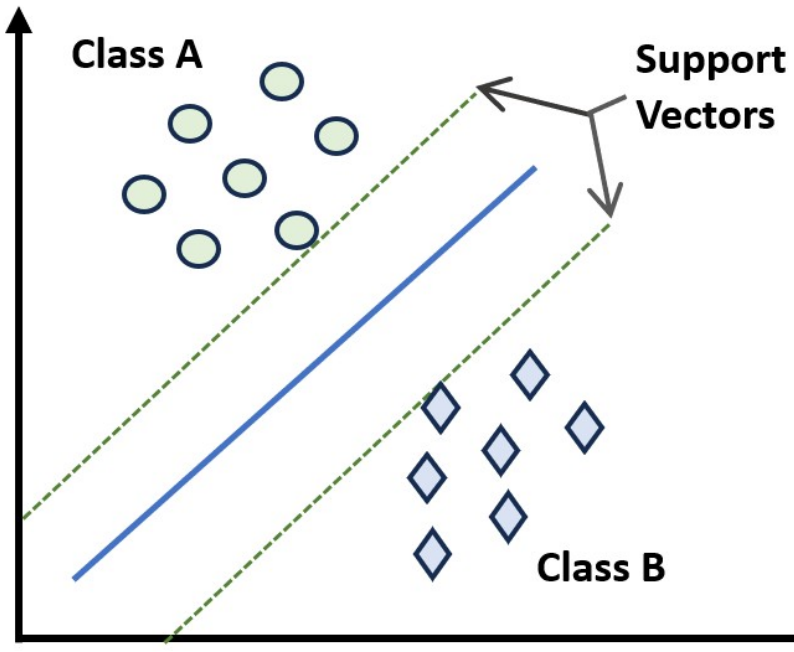

Figure 5. Support Vector Machine. This classifier tries to find the best possible hyperplane that separates data points into different classes.

Maximizing the margin leads to better generalization performance. A more considerable margin means the model is less sensitive to noise and outliers, making it more robust. The Scikit-learn library [33] was used to implement this classifier. The kernel, C, and gamma values were 'rbf', 1.0, and 'scale', respectively.

3.5.2. Logistic Regression

This work also applied a logistic regression model as a classification method. This model is a supervised machine learning algorithm used for classification tasks. The goal is to predict the probability that an element belongs to a given class [34]. Logistic regression is used for binary classification, where we use the sigmoid function (Equation (1)), which produces a probability value between 0 and 1.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

For example, we have classes Class 0 (non-cognitive impairment) and Class 1 (cognitive impairment). If the value of the logistic function for an input is more significant than 0.5, then it belongs to Class 1. On the other hand, if we obtain any other value, it belongs to Class 0. This algorithm is an extension of linear regression but is mainly used for classification problems. We used the logistic regression model from the SKLearn Python package (Python 3.11.13 and SkLearn 1.6.1) [33]. We set the 'C' and 'solver' hyperparameters to 1 and 'liblinear', respectively.

4. Results

The analysis of the characteristics of the samples of subjects with cognitive impairment and without cognitive impairment, considering the repetition of ideas, use of circumlocutions, illogical flow, and incomplete ideas, generated the results shown in Figure 6.

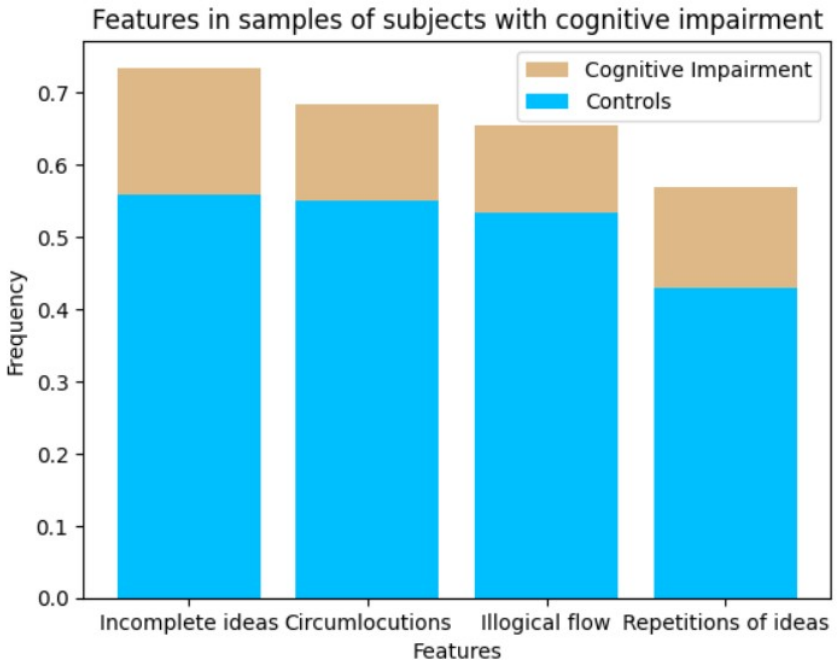

Figure 6. Characteristics analyzed in the sample group of subjects with and without cognitive impairment. Incomplete ideas were the most important feature in the samples of cognitive impairment subjects.

Figure 6 illustrates that incomplete ideas were the most predominant feature in the samples of subjects with cognitive impairment, while the repetition of ideas was the least predominant feature.

However, to test the proposed methodology, the following commonly used metrics in similar studies were utilized: accuracy (Equation (2)), precision (Equation (3)), recall (Equation (4)), and F1-score (Equation (5)). These metrics are defined in terms of the correctly estimated values, namely true positives (TPs) and true negatives (TNs), as well as the incorrect predictions, namely false positives (FPs) and false negatives (FNs).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{5}$$

K-fold cross-validation was also employed. A value of k = 5 was used to test the effectiveness of the data augmentation method. The data were divided into five folds, split into training and testing, using 80% of the data for training and 20% for testing.

Table 2 shows the results of comparing different embeddings models using an SVM as a classifier.

Table 2. The results of the proposed method are based on four different embedding models.

Model	Accuracy	Precision	Recall	F1-Score
RoBERTa	**0.8650**	**0.8659**	**0.86525**	**0.8649**
BERT	0.85111	0.855246	0.851367	0.8507
Linq-Embed-Mistral	0.8491249	0.856885	0.8491249	0.8482079
gte-Qwen2-7B-instruct	0.845249	0.8507056	0.845249	0.84456

To compare our proposed methodology, four data augmentation techniques were used. Three were based on [35] (Synonym Replacement, Random Swap, and Random Deletion), and one was based on word replacement, considering BERT embeddings. These techniques were implemented using the nlpaug library.

Tables 3 and 4 show the results of each metric for the proposed method and the other text augmentation techniques using the SVM and logistic regression classifier, respectively. The reported results are the average of 10 iterations.

Table 3. Results of the proposed method compared with other text augmentation techniques using the SVM model as a classifier and RoBERTa as an embedding model.

Model	Accuracy	Precision	Recall	F1-Score
Proposed method	**0.8742** **	**0.8736** *	**0.8781** **	**0.8744** **
Synonym replacement	0.8629	0.8626	0.8611	0.8642
Random Swap	0.8459	0.8446	0.8481	0.8442
Random Deletion	0.8462	0.8466	0.8481	0.8462
BERT embeddings	0.8580	0.8596	0.8571	0.8568
No data augmentation	0.8650	0.8659	0.86525	0.8649

Note: Statistically significant improvements are indicated by * ($p < 0.01$) and ** ($p < 0.001$). The results were obtained using the Mann–Whitney test.

Table 4. Results of the proposed method compared with other text augmentation techniques using the logistic regression model as a classifier and RoBERTa as an embedding model.

Model	Accuracy	Precision	Recall	F1-Score
Proposed method	**0.8462**	**0.8456**	**0.8474**	**0.8463**
Synonym replacement	0.8335	0.8326	0.8337	0.8329
Random Swap	0.827	0.8246	0.8281	0.8262
Random Deletion	0.8225	0.8236	0.8211	0.8218
BERT embeddings	0.8402	0.8406	0.8511	0.8408
No data augmentation	0.8401	0.8399	0.8409	0.8404

The results showed that the proposed model outperforms other text data augmentation methods.

Figure 7 presents the accuracy results for the ten experiments performed with the five data augmentation techniques using box plots and the SVM model as a classifier.

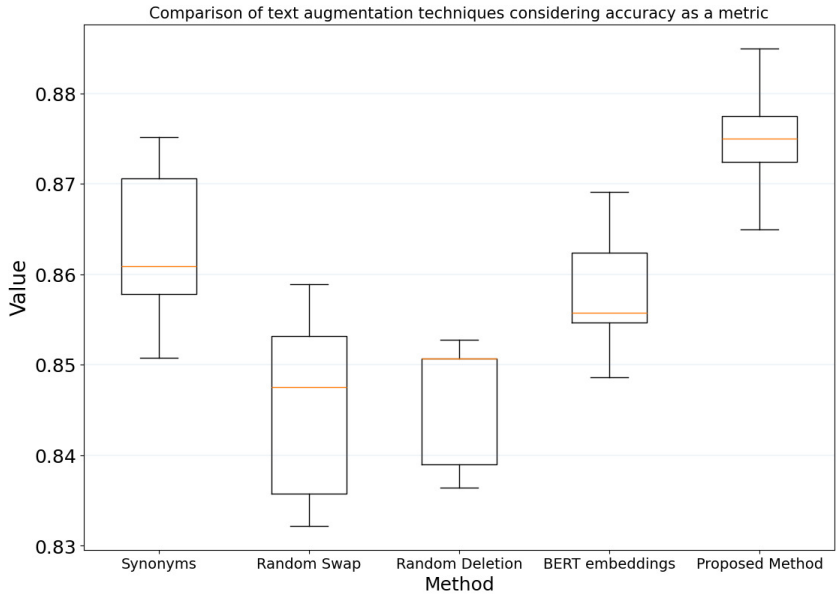

Figure 7. Box plots of the proposed method compared to traditional data augmentation methods using accuracy as a metric.

The box plots shown in Figure 7 indicate that the proposed method and the BERT-based synonyms showed the best performance in terms of variance. Furthermore, the Random Swap method yielded the lowest accuracy and exhibited higher variance than the other methods.

To increase the model's interpretability, an attention mechanism was developed to analyze each sentence in the text and observe whether there were changes in the attention score before and after data augmentation. For this model, a classifier was developed using an attention mechanism [14] for each sentence (Figure 8).

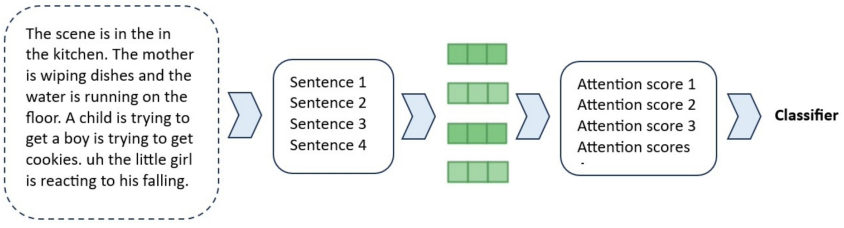

Figure 8. Interpretability model using attention scores for each sentence.

As shown in Figure 9, when data augmentation is performed for the example where the characteristic of idea repetition exists, a change in the attention score is observed, which helps identify the characteristics the model is analyzing to make the prediction.

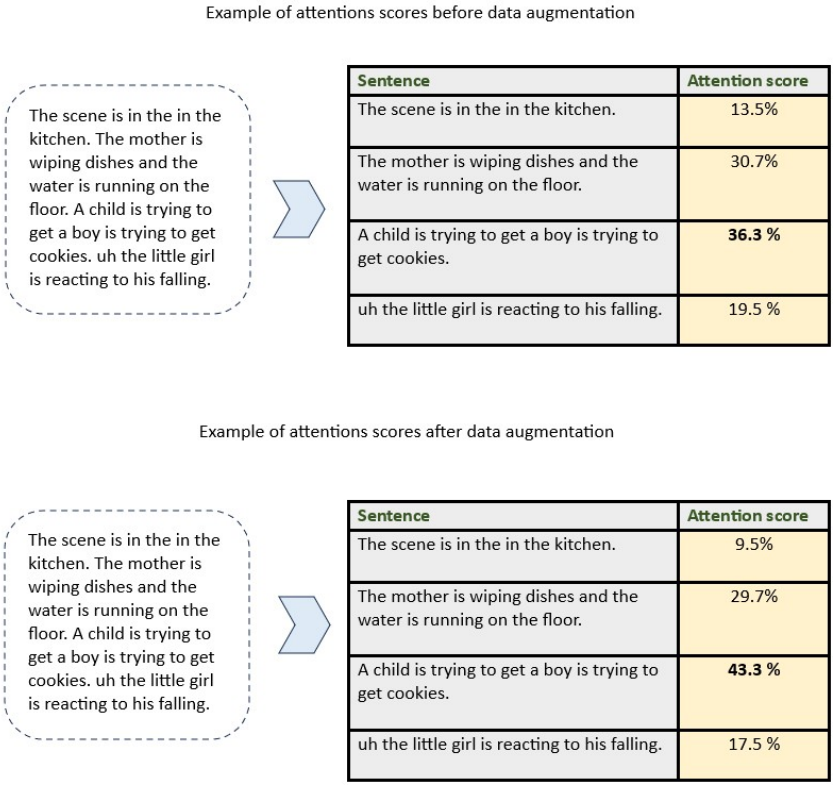

Figure 9. Example of a change in the attention score on a sentence with idea repetition before and after data augmentation.

5. Discussion

The results obtained from both classification models indicate that leveraging common symptoms of cognitive impairment within a text augmentation framework substantially improves classification task outcomes. Based on the rigorous evaluation using the proposed metrics, this methodology presents a robust and effective approach for data augmentation in the context of neuropsychological assessments.

Regarding the embedding models used, the results showed that the RoBERTa model demonstrated the best performance, even outperforming more advanced models such as Linq-Embed-Mistral and gte-Qwen2-7B-instruct, which may indicate that simpler models may have comparable results in classification tasks involving small texts with a reduced language with more complex models. Likewise, these results could be due to differences in the models' pre-training.

The synonym replacement method was the best traditional technique. The possible reason is that, as patients identified as having AD, word retrieval is the most notable characteristic in oral descriptions [36,37]. This can lead to using synonyms for some words because they cannot remember the correct word. Furthermore, word retrieval deficits, measured using verbal fluency tasks, are hallmark features of the early clinical stages of AD [38].

The synonym replacement model using BERT's contextual word embeddings and the proposed method showed the best results regarding variance. This may indicate that the modifications to the samples were minor compared to those performed by the other text augmentation methods. This performance suggests that in the case of word

replacement using BERT embeddings, the modifications to the texts were less aggressive than those performed by other methods, such as random word deletion. In the case of the proposed method, because only the number of samples that met the conditions for each group was increased, better metrics and lower variance were obtained compared to traditional methods.

Compared to similar studies, Kang et al. [24] propose generating synthetic data from clinical interviews to detect traits associated with depression. This model demonstrates the utility of thought chains in generating synthetic data in clinical settings. Due to this fact, the proposed model incorporates the DeepSeek Reasoner model to analyze the characteristics of the samples to be augmented, thereby obtaining an in-depth analysis of the possible signals in the subjects' responses. On the other hand, the proposed framework of Wu et al. [39] generated synthetic data using LLMs for the mental health domain, focusing on tasks such as suicidal ideation detection, which highlights the potential and ethical considerations of using LLMs for sensitive clinical narratives. Moreover, the work highlights the importance of symptomatic analysis for generating feature data, as in our proposed work.

A potential problem with the proposed model is the overfitting of the classifier due to the synthetically generated semantic and pragmatic features, which could increase the risk of architectural overfitting, as studied by Pratticò et al. [40].

6. Conclusions

Analyzing linguistic features in patients with cognitive impairment presents a significant challenge due to the condition's inherent complexity. Nevertheless, the proposed data augmentation method, derived from observed linguistic characteristics in individuals with this syndrome, offers a viable alternative to conventional artificial intelligence models in resource-constrained environments. Unlike generic augmentation techniques, this method reflects genuine clinical linguistic patterns more accurately.

This study introduces a novel analytical approach to investigating cognitive impairment by incorporating complex linguistic features, such as circumlocutions. Recent advancements in reasoning models have enabled the effective analysis of these intricate features within subject samples.

In critical healthcare contexts, where clinical decisions directly impact patient well-being, the interpretability of an artificial intelligence (AI) model's output is as crucial as its predictive accuracy. Understanding the mechanistic basis of an AI-driven prediction or diagnosis is essential for clinical adoption. Therefore, the proposed model's integrated explainability facilitates its seamless integration into clinical workflows, thereby fostering greater confidence among healthcare professionals.

One disadvantage of the embedding models analyzed, especially the latest state-of-the-art ones, such as Linq-Embed-Mistral and gte-Qwen2-7B-instruct, is that they are challenging to integrate in practical clinical settings with limited resources due to computational cost and time complexity. A BERT-based model or a TF-IDF model could be better options in these scenarios. Specifically, the latter is a lightweight option that presents comparable performance with state-of-the-art models on the cookie theft description task Llaca-Sánchez et al. [41].

A limitation of this study is its narrow scope, as only four characteristics associated with cognitive impairment were considered. Future research could incorporate a broader range of characteristics to develop a more robust model for analyzing language in individuals with cognitive impairment.

Future investigations could explore additional features for sample analysis and alternative strategies for sample selection and classification. This includes generating samples with diverse characteristics and evaluating various subject classification methodologies,

as well as exploring other vectorization techniques, such as TF-IDF, which have provided interesting metrics in the analyzed test Llaca-Sánchez et al. [41]. In addition, we aim to integrate other characteristics, such as prosodic flattening or lexical disfluency, as well as to include language characteristics of subjects in moderate and advanced stages of the disease. In addition, we aim to integrate other characteristics, such as prosodic flattening or lexical disfluency, as well as language characteristics of subjects in moderate and advanced stages of the disease. Moreover, we aim to test the proposed methodology on datasets in other languages, such as the one used by Díaz-Rivera et al. [42], to evaluate its applicability in other populations.

Likewise, another future work is to analyze the method when the subject has a co-morbid condition, for example, when AD and Parkinson's disease or AD and depression are present. In the latter case, it has been studied that both depression and Alzheimer's disease present in different syndromes; therefore, as mentioned by Fraser et al. [43], it is probably unrealistic to delineate the many possible combinations of depression, AD, and other possible medical conditions by analyzing only a single language task.

Because the purpose of this study is to develop a model for facilitating the early diagnosis of cognitive impairment, a mobile application is expected to be developed, which will analyze both audio and transcripts for indicators of cognitive decline. This will serve as a helpful tool for specialists. For now, the proposed method is limited to text only, but future work is expected to contribute to analyzing audio for indicators of cognitive decline. Additionally, integrating multiple signal modalities [44] and combining them with acoustic prosody would further enrich this study.

Author Contributions: Conceptualization, L.R.G.-N. and S.T.-A.; methodology, L.R.G.-N., S.S.-C. and S.M.-R.; writing—original draft preparation, L.R.G.-N., S.S.-C. and S.T.-A.; writing—review and editing, L.R.G.-N., S.S.-C., S.T.-A. and S.M.-R.; supervision, S.T.-A., S.S.-C. and S.M.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code used in this study is freely available at GitHub: https://github.com/luisroberto-maker/DataAugDementiaPaper, accessed on 4 August 2025.

Acknowledgments: The authors wish to thank DementiaBank for providing the data.

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

References

1. Leong, D.P.; Teo, K.K.; Rangarajan, S.; Lopez-Jaramillo, P.; Avezum, A.; Orlandini, A.; Seron, P.; Ahmed, S.H.; Rosengren, A.; Kelishadi, R.; et al. Prognostic value of grip strength: Findings from the Prospective Urban Rural Epidemiology (PURE) study. *Lancet* **2015**, *386*, 266–273. [CrossRef]
2. Penfold, R.B.; Carrell, D.S.; Cronkite, D.J.; Pabiniak, C.; Dodd, T.; Glass, A.M.; Johnson, E.; Thompson, E.; Arrighi, H.M.; Stang, P.E. Development of a machine learning model to predict mild cognitive impairment using natural language processing in the absence of screening. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 129. [CrossRef] [PubMed]
3. Arriba-Pérez, F.; García-Méndez, S.; González-Castaño, F.J.; Costa-Montenegro, E. Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities. *J. Ambient Intell. Humaniz. Comput.* **2022**, *14*, 16283–16298. [CrossRef] [PubMed]
4. Croot, K.; Hodges, J.R.; Xuereb, J.; Patterson, K. Phonological and articulatory impairment in Alzheimer's disease: A case series. *Brain Lang.* **2000**, *75*, 277–309. [CrossRef]
5. Mascio, A.; Stewart, R.; Botelle, R.; Williams, M.; Mirza, L.; Patel, R.; Pollak, T.; Dobson, R.; Roberts, A. Cognitive impairments in schizophrenia: A study in a large clinical sample using natural language processing. *Front. Digit. Health* **2021**, *3*, 711941. [CrossRef] [PubMed]

6. Rosas, D.S.; Arriaga, S.T.; Fernández, M.A.A. Search for dementia patterns in transcribed conversations using natural language processing. In Proceedings of the 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico City, Mexico, 11–13 September 2019; pp. 1–6.
7. Santander-Cruz, Y.; Salazar-Colores, S.; Paredes-García, W.J.; Guendulain-Arenas, H.; Tovar-Arriaga, S. Semantic feature extraction using SBERT for dementia detection. *Brain Sci.* **2022**, *12*, 270. [CrossRef]
8. Ramos-Morales, I.D. Extraction of Dementia Features from Audio and Text Records Using Machine Learning Algorithms. 2025. Available online: https://ri-ng.uaq.mx/bitstream/123456789/11297/1/IGMAC-317932.pdf (accessed on 4 August 2025).
9. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
10. Bayer, M.; Kaufhold, M.A.; Reuter, C. A survey on data augmentation for text classification. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]
11. 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]
12. Srivastava, S.K.; Singh, S.K.; Suri, J.S. A healthcare text classification system and its performance evaluation: A source of better intelligence by characterizing healthcare text. In *Cognitive Informatics, Computer Modelling, and Cognitive Science*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 319–369.
13. Masrani, V.; Murray, G.; Field, T.; Carenini, G. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In Proceedings of the BioNLP 2017, Vancouver, BC, Canada, 4 August 2017; pp. 232–237.
14. Chen, J.; Zhu, J.; Ye, J. An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 4085–4089.
15. Lin, K.; Washington, P.Y. Multimodal deep learning for dementia classification using text and audio. *Sci. Rep.* **2024**, *14*, 13887. [CrossRef]
16. Mueller, K.D.; Hermann, B.; Mecollari, J.; Turkstra, L.S. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *J. Clin. Exp. Neuropsychol.* **2018**, *40*, 917–939. [CrossRef]
17. Goodglass, H.; Kaplan, E. *Boston Diagnostic Aphasia Examination Booklet*; Lea & Febiger: Philadelphia, PA, USA, 1983.
18. MacWhinney, B. Language Emergence. 2002. Available online: https://www.taylorfrancis.com/books/mono/10.4324/9781315805672/childes-project-brian-macwhinney (accessed on 4 August 2025).
19. Jaramillo, J.H. Demencias: Los problemas de lenguaje como hallazgos tempranos. *Acta Neurol. Colomb.* **2010**, *26*, 101–111.
20. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.
21. Zhan, Z.; Zhou, S.; Zhou, H.; Deng, J.; Hou, Y.; Yeung, J.; Zhang, R. An evaluation of deepseek models in biomedical natural language processing. *arXiv* **2025**, arXiv:2503.00624.
22. Lira, J.O.d.; Minett, T.S.C.; Bertolucci, P.H.F.; Ortiz, K.Z. Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease. *Dement. Neuropsychol.* **2014**, *8*, 260–265. [CrossRef]
23. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
24. Kang, A.; Chen, J.Y.; Lee-Youngzie, Z.; Fu, S. Synthetic data generation with LLM for improved depression prediction. *arXiv* **2024**, arXiv:2411.17672. [CrossRef]
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
26. 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.
27. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]
28. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. Qwen2 Technical Report. *arXiv* **2024**, arXiv:2407.10671.
29. Gweon, H.; Schonlau, M. Automated classification for open-ended questions with BERT. *J. Surv. Stat. Methodol.* **2024**, *12*, 493–504. [CrossRef]
30. Muennighoff, N.; Tazi, N.; Magne, L.; Reimers, N. MTEB: Massive text embedding benchmark. *arXiv* **2022**, arXiv:2210.07316.
31. Shakeri, A.; Freja, S.A.; Hallaj, Y.; Farmanbar, M. Uncovering Linguistic Patterns: A Machine Learning Exploration for Early Dementia Detection in Speech Transcripts. In Proceedings of the 2024 4th International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway, 16 April 2024; pp. 1–8.
32. Shakeri, A.; Farmanbar, M. Natural language processing in Alzheimer's disease research: Systematic review of methods, data, and efficacy. *Alzheimer's Dement. Diagn. Assess. Dis. Monit.* **2025**, *17*, e70082. [CrossRef]

33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
34. Geng, Y.; Li, Q.; Yang, G.; Qiu, W. Logistic regression. In *Practical Machine Learning Illustrated with KNIME*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 99–132.
35. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196. [CrossRef]
36. Croisile, B.; Ska, B.; Brabant, M.J.; Duchene, A.; Lepage, Y.; Aimard, G.; Trillet, M. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain Lang.* **1996**, *53*, 1–19. [CrossRef] [PubMed]
37. Kavé, G.; Goral, M. Word retrieval in connected speech in Alzheimer's disease: A review with meta-analyses. *Aphasiology* **2018**, *32*, 4–26. [CrossRef]
38. Putcha, D.; Dickerson, B.C.; Brickhouse, M.; Johnson, K.A.; Sperling, R.A.; Papp, K.V. Word retrieval across the biomarker-confirmed Alzheimer's disease syndromic spectrum. *Neuropsychologia* **2020**, *140*, 107391. [CrossRef]
39. Wu, Y.; Mao, K.; Zhang, Y.; Chen, J. CALLM: Enhancing clinical interview analysis through data augmentation with large language models. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 7531–7542. [CrossRef]
40. Pratticò, D.; Laganà, F.; Oliva, G.; Fiorillo, A.S.; Pullano, S.A.; Calcagno, S.; De Carlo, D.; La Foresta, F. Integration of LSTM and U-Net models for monitoring electrical absorption with a system of sensors and electronic circuits. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2533311. [CrossRef]
41. Llaca-Sánchez, B.A.; García-Noguez, L.R.; Aceves-Fernández, M.A.; Takacs, A.; Tovar-Arriaga, S. Exploring LLM Embedding Potential for Dementia Detection Using Audio Transcripts. *Eng* **2025**, *6*, 163. [CrossRef]
42. Díaz-Rivera, M.N.; Birba, A.; Fittipaldi, S.; Mola, D.; Morera, Y.; de Vega, M.; Moguilner, S.; Lillo, P.; Slachevsky, A.; González Campo, C.; et al. Multidimensional inhibitory signatures of sentential negation in behavioral variant frontotemporal dementia. *Cereb. Cortex* **2023**, *33*, 403–420. [CrossRef]
43. Fraser, K.C.; Rudzicz, F.; Hirst, G. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, San Diego, CA, USA, 16 June 2016; pp. 1–11.
44. Laganà, F.; Pratticò, D.; Angiulli, G.; Oliva, G.; Pullano, S.A.; Versaci, M.; La Foresta, F. Development of an Integrated System of sEMG Signal Acquisition, Processing, and Analysis with AI Techniques. *Signals* **2024**, *5*, 476–493. [CrossRef]