Multimodal Classification of Dementia: Fine-tuning Wav2vec and Word2vec Using Audio, Text, and Timestamps

Kaiying Lin KYLIN@HAWAII.EDU
University of Hawai'i, Mānoa, HI, USA

Peter Y. Washington PYW@HAWAII.EDU
University of Hawai'i, Mānoa, HI, USA

Abstract

Dementia is a progressive neurological disorder that profoundly affects the daily lives of older adults, impairing abilities such as verbal communication and cognitive function. Early diagnosis is essential for enhancing both lifespan and quality of life for affected individuals. Despite its importance, diagnosing dementia is complex and often necessitates a multimodal approach incorporating diverse clinical data types. In this study, we fine-tune Wav2vec and Word2vec baseline models using two distinct data types: audio recordings and text transcripts. We experiment with four conditions: original datasets versus datasets purged of short sentences, each with and without data augmentation. Our results indicate that synonym-based text data augmentation generally enhances model performance, underscoring the importance of data volume for achieving generalizable performance. Additionally, models trained on text data frequently excel and can further improve the performance of other modalities when combined. The selection and integration of data modalities are crucial factors influencing the effectiveness of dementia screening models.

Keywords: dementia, multimodality, fine-tuning, wav2vec, word2vec, timestamps

1. Introduction

Dementia is a complex syndrome characterized by a decline in cognitive functions such as memory, thinking, and reasoning. An estimated 47.5 million people worldwide are affected by dementia, with some experiencing severe emotional and language impairments. Recognizing these serious consequences underscores the critical importance of early diagnosis in clinical practice.

The diagnostic process for dementia often includes a comprehensive review of the patient's medical history, genetic testing, psychiatric evaluations, and cognitive assessments, often supplemented by neuroimaging techniques. Given the multi-faceted nature of dementia diagnosis, there is growing interest in streamlining the process through scalable, accessible, and low-cost methods. Among the cognitive deteriorations caused by dementia, verbal and speech impairments are notable, making verbal fluency a promising early diagnostic indicator.

One widely-used assessment tool for verbal fluency is the Verbal Fluency Test (VFT), which measures both the speed and thematic organization of word production. The Dementia Databank (Lanzi et al., 2023), the largest publicly available dataset on dementia, offers data derived from patients undergoing these assessments. These datasets, consisting of audio recordings and text transcripts, provide a rich resource for machine learning (ML) models aimed at distinguishing between healthy individuals and those with dementia.

Previous research efforts have employed machine learning models for dementia detection, with some studies fine-tuning pre-existing language models(e.g., (Yuan et al., 2020), and others developing models from the ground up (e.g., (Luz et al., 2020). However, most prior work has focused on singular data types—either audio (e.g., (Torre et al., 2021; Chlasta and Wołk, 2021)) or text data (e.g., (Guo et al., 2021))—for model training. Fewer studies have explored the synergistic effects of integrating these different data types(e.g., (Sarawgi et al., 2020; Hlédiková et al., 2022)).

In this paper, we leverage multiple data sources within a single model, incorporating audio, text, and timestamps from the Dementia Databank, to perform a comprehensive analysis for dementia detec-

tion. Utilizing pre-trained embeddings from Wav2vec and Word2vec, we aim to enhance the efficiency and effectiveness of diagnostic models. The paper is structured as follows: Section 2 provides a brief review of related work; Section 3 introduces six distinct models, each featuring different combinations of data modalities along with our data augmentation techniques; Section 4 delves into the details of our experimental methodology; Section 5 presents the findings from our experiments; and Section 6 offers concluding remarks. To the best of our knowledge, this is the first study to integrate time embeddings with text and audio data for multimodal dementia diagnosis.

2. Related work

Many previous research efforts have focused primarily on detecting a specific type of dementia, such as Alzheimer's Disease (AD). Within the Dementia Databank, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge (Luz et al., 2020) offers multiple shared tasks, allowing researchers to base their methodologies on common datasets for comparative analysis. Prior AD detection techniques in these shared tasks have employed fine-tuning of existing models, data augmentation, and feature engineering. Studies that utilize feature engineering (Luz et al., 2020; Balagopalan et al., 2020; Sarawgi et al., 2020; Chlasta and Wołk, 2021) have extracted features—either manually or through existing packages—and trained models on binary classification tasks. Other research efforts have fine-tuned pre-trained language models like BERT (Devlin et al., 2019) to achieve similar goals (Balagopalan et al., 2020; Guo et al., 2021). Data augmentation strategies have also been applied to mitigate the challenge of limited data availability (Hlédiková et al., 2022).

In addition to aiming for high performance in detection tasks, an important objective is the identification of features that can assist with AD diagnosis in clinical settings. For instance, (Balagopalan et al., 2020) highlighted various semantic and lexico-syntactic features, such as the proportion of personal pronouns and average sentence length.

Beyond the ADReSS Challenge, researchers have also explored the Pitt Corpus (Becker et al., 1994) within the Dementia Databank. Some studies have constructed models from scratch, introducing various model variations (e.g., Karlekar et al., 2018), while others have leveraged pre-existing models (e.g., Matošević and Jović, 2022). Among these, some studies have solely used text transcripts (e.g., Guo et al., 2020), while others have focused exclusively on audio recordings (e.g., Guo et al., 2021). Only a few have integrated multiple modalities, incorporating both audio and text data for dementia detection (e.g., Ilias et al., 2022; Sarawgi et al., 2020).

Relying solely on one data type can compromise the effectiveness of a dementia diagnosis, given that a comprehensive diagnosis typically requires insights from multiple data sources. Models that integrate multiple data types could offer more robust and efficient diagnostic capabilities.

3. Methodology

We evaluate two data modalities, audio and text, as well as text-based synonym data augmentation.

3.1. Audio model

We created an audio model which was fine-tuned using Wav2vec as the baseline representation. The audio data was processed through Wav2vec to obtain audio embeddings, which were then subjected to a dense layer for binary classification, using binary cross-entropy loss as the evaluation metric. Note that the weights from the pretrained Wav2vec was frozen during the training and only the other layers of architecture were updated.

3.1.1. Wav2vec

Wav2vec (Baevski et al., 2020) is a self-supervised convolutional architecture that transforms audio waveforms into representative embeddings. Initially trained on unlabeled audio data, these embeddings are then processed through a transformer for a masked task. In this task, half of the audio embeddings are masked and predicted by the remaining unmasked portions. Wav2vec is particularly useful in speech recognition tasks due to its adaptability to various audio recordings and its inherently robust performance.

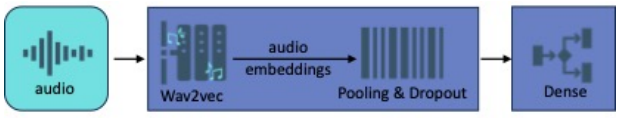


Figure 1: Architecture of Audio Model

3.2. Text model

The text model incorporates the embedding layers from Word2vec and utilizes an LSTM model connected to a dense layer for classification tasks.

3.2.1. Word2vec

Word2vec (Mikolov et al., 2013) is a feed-forward neural network designed to produce vector representations of words. It uses surrounding words as input to generate these vectors, capturing semantic relationships between the words. The resulting vectors effectively position semantically similar words closer in the vector space. Again, similar to the audio model, the weights from the pretrained Word2vec was frozen during the training and only the other layers of architecture were updated.

3.2.2. LSTM

We employed an LSTM model with 16 units to process embedded sentences and used a dropout and recurrent dropout rate of 0.2. A dense layer with sigmoid activation is appended to the LSTM layer to perform binary classification.

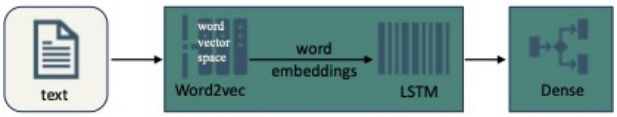


Figure 2: Architecture of Text Model

3.3. Timestamps

Timestamps for each word were extracted from the corpus. In models that combined text and time (Figure 3), these timestamps were concatenated with the word embeddings before feeding them as input into subsequent layers. In models incorporating audio and time (Figure 4), timestamps were processed through an LSTM layer, following the extraction of audio embeddings, which were then passed through an average pooling layer and a dropout layer prior to classification.

3.4. Concatenated model

In the concatenated audio-text model (Figure 5), word embeddings from the text model were processed through an LSTM layer. Both the audio and text models were then passed through the same average pooling and dropout layers before their concatenation. A final dense layer was added for classification tasks. We also developed a model combining audio, text, and timestamps (Figure 6). The architecture for text and timestamps remained consistent with their individual models, as did the audio model. These were then concatenated for the final classification task.

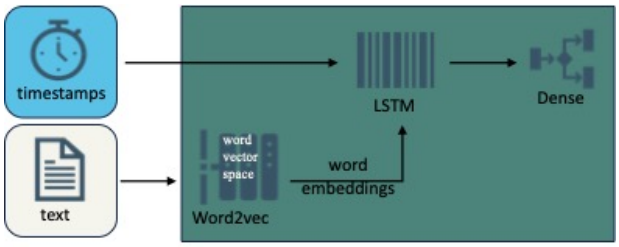


Figure 3: Architecture of Text+Timestamps Model

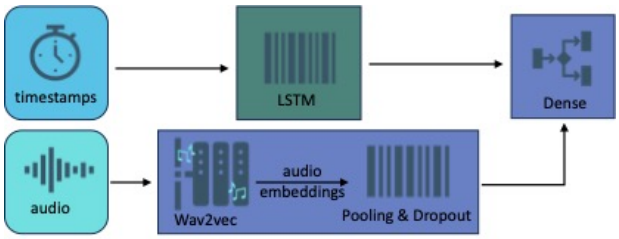


Figure 4: Architecture of Audio+Timestamps Model

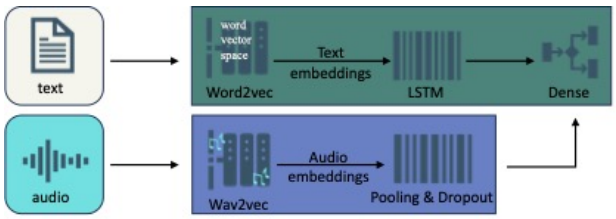


Figure 5: Architecture of Audio+Text Model

3.5. Data augmentation

Due to the lack of a substantial volume of data points in the original dataset, we implemented data augmentation techniques. Specifically, we employed the

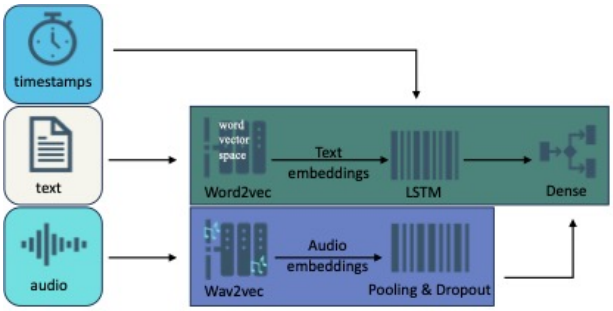


Figure 6: Architecture of Audio+Text+Timestamps Model

Synonym Replacement (SR) method (Wei and Zou, 2019) where a synonym for a word is identified and used to create a duplicated sentence with the original word replaced by its synonym. Each word was replaced by its synonym twice (n=2).

4. Experiments

4.1. Datasets and Data Preprocessing

Dataset source Our dataset comes from the "Pitt Cookie Theft" folder of the Dementia Databank (Becker et al., 1994). This dataset contains participants' responses when asked to describe what they see in stimulus photographs. Unique to this dataset in the Databank is the inclusion of timestamps for each word, which enabled us to use it for our multimodal models.

Data Preparation Both audio and text data were divided into individual sentences, with each sentence serving as a data point for training. The dataset contains 9,447 data points, of which 3,873 are dementia-related and 5,574 are control points. These numbers were derived by categorizing the sentences spoken by investigators as control, as well as those spoken by patients in the control dataset from the Cookie folder. Dementia datapoints, conversely, only include sentences spoken by patients classified as dementia cases in the same folder.

For audio preprocessing, the dataset was first processed through a Wav2vec feature extractor, ensuring compatibility with the sampling rates used during the model's pre-training. The text data underwent tokenization using a custom dictionary, mapping words to their corresponding pre-trained word2vec embeddings in Gensim's 'word2vec-google-news-300' (Rehurek and Sojka, 2011). Words without corresponding embeddings were marked as Out-of-Vocabulary (OOV) and were represented by zero vectors.

Timestamps for each word were retained, indicating the starting and ending times. The timestamp for the first word in each sentence was normalized to begin at 0 and processed as decimal digits.

Four different dataset conditions were created:

Original Condition: The original dataset with 9,447 datapoints, including 3,873 dementia and 5,574 control datapoints.

Shorts-Removed Condition: Excluded sentences shorter than two words, resulting in 4,318 control and 3,368 dementia datapoints.

Original-Augmented Condition: Augmented from the dataset in Original Condition, leading to 31,273 control and 22,664 dementia datapoints.

Shorts-Augmented Condition: Augmented from the dataaset in Short-Removed Condition, yielding 28,964 control and 22,039 dementia datapoints.

For all conditions, the datasets were divided into training and test sets using a 4:1 ratio. Furthermore, the training datasets were also split into training and validation sets with a 4:1 ratio, facilitating 5-fold cross-validation for hyperparameter optimization.

4.2. Experimental Setup

All models were trained for 50 epochs with a batch size of 16. The objective was to minimize binary cross-entropy loss. To prevent overfitting, early stopping was implemented, which halted training if the validation loss failed to decrease for 10 consecutive epochs. The implementation was carried out using the TensorFlow Keras library (Chollet et al., 2015).

4.3. Results

The experiment was conducted with five separate and independent train-validation splits to ensure generalizability and reliability. The results were then averaged, and both the mean and standard deviation were reported. Evaluation metrics included: accuracy, precision, recall, F1 score, and AUC ROC scores. The highest validation scores for each metric are reported in bold.

Our results highlight the challenges and opportunities associated with utilizing multimodal data for dementia detection. Specifically, as evidenced in Table 1, Figures 7a and 8a, unimodal audio models underperformed compared to their text counterparts.

Table 1: Results from Original Condition

Original	Accuracy	Precision	Recall	F1-score	AUROC
Audio	0.6484±0.008	0.593±0.019	0.4425±0.063	0.5039±0.032	0.7085±0.011
Text	**0.691±0.034**	0.6484±0.07	0.6299±0.127	0.62765±0.027	0.7638±0.046
Audio+Time	0.6123±0.006	0.5565±0.022	0.3286±0.039	0.4115±0.026	0.6517±0.011
Text+Time	0.6909±0.013	0.6537±0.036	0.5566±0.4463	0.5995±0.022	**0.7647±0.018**
Audio+Text	0.6731±0.04	0.5958±0.047	**0.6852±0.097**	**0.6341±0.048**	0.7448±0.045
Audio+Text+Time	0.6539±0.031	0.5874±0.07	0.5301±0.138	0.55±0.087	0.7161±0.043

Table 2: Results from Original-augmented Condition

Original-augmented	Accuracy	Precision	Recall	F1-score	AUROC
Audio	0.6038±0.036	0.5846±0.021	0.1831±0.04	0.2764±0.044	0.6336±0.003
Text	0.8294±0.005	**0.8339±0.022**	0.751±0.042	0.7892±0.013	0.9208±0.002
Audio+Time	0.6306±0.068	0.5994±0.012	0.3673±0.044	0.4542±0.033	0.6759±0.008
Text+Time	**0.8344±0.006**	0.8267±0.035	0.7721±0.043	**0.797±0.009**	**0.9236±0.005**
Audio+Text	0.825±0.013	0.7978±0.039	**0.7859±0.056**	0.7899±0.019	0.9124±0.015
Audio+Text+Time	0.8315±0.014	0.8212±0.034	0.767±0.012	0.7927±0.014	0.9177±0.014

Table 3: Results from Shorts-removed Condition

Shorts-removed	Accuracy	Precision	Recall	F1-score	AUROC
Audio	0.5958±0.014	0.587±0.046	0.2945±0.108	0.3811±0.084	0.624±0.022
Text	**0.6861±0.027**	**0.623±0.042**	0.6951±0.058	**0.6545±0.022**	**0.7593±0.024**
Audio+Time	0.5897±0.009	0.5631±0.034	0.3554±0.07	0.4306±0.054	0.6163±0.01
Text+Time	0.6683±0.029	0.6197±0.043	**0.7011±0.137**	0.6494±0.032	0.7353±0.047
Audio+Text	0.6052±0.014	0.5669±0.024	0.5255±0.152	0.534±0.082	0.6482±0.014
Audio+Text+Time	0.6257±0.065	0.5724±0.075	0.5239±0.151	0.5415±0.107	0.6769±0.081

Table 4: Results from Shorts-augmented Condition

Shorts-augmented	Accuracy	Precision	Recall	F1-score	AUROC
Audio	0.5954±0.001	0.5989±0.025	0.1939±0.029	0.2914±0.03	0.6258±0.008
Text	0.841±0.01	0.8212±0.014	**0.8089±0.024**	0.8148±0.014	0.9276±0.009
Audio+Time	0.6254±0.005	0.5894±0.023	0.4298±0.05	0.4951±0.032	0.6692±0.008
Text+Time	**0.8478±0.003**	**0.8375±0.02**	0.8039±0.207	**0.8199±0.007**	**0.9345±0.005**
Audio+Text	0.835±0.02	0.8154±0.036	0.7982±0.039	0.80591±0.023	0.9216±0.02
Audio+Text+Time	0.8451±0.003	0.83646±0.027	0.7992±0.028	0.8166±0.04	0.931±0.005

The audio+time model as in Figures 7a and 8d also yielded suboptimal results, further highlighting that audio, as a modality, may be challenging to engineer a useful feature representation for, with current state-of-the-art methods such as Wav2vec not providing enough specificity and expressiveness of the relevant features. On the other hand, the text model (Table 1, Figures 7a and 8b) excelled across the board, and its efficacy was only augmented when coupled with time, as demonstrated by the superior performance of the text+time model (Table 1, Figures 7a and 8e).

We observed higher standard deviations in some modalities, primarily in the audio-based models. This suggests that the model was more prone to poor fitting in several data splits.

As observed in Table 3 and Figures 7c, the exclusion of shorter sentences during preprocessing did not bring significant improvement in the overall model performance. This indicates that the initial preprocessing strategy was robust and captured the most important aspects of the data for the task at hand.

Table 2, Figures 7b and Table 4, Figure 7d illustrate the significant improvement achieved as a result of our data augmentation techniques. AUROC scores in models incorporating text data surpassed 90% Figure 8b, 8c, 8d and 8e), and both accuracy and F1 scores were consistently above 80%. This uplift in performance metrics strongly suggests that the augmented data captured richer semantic and syntactic features essential for dementia detection.

The relative consistency in high performance across metrics for text-based models further emphasizes the influence that the textual modality has on the overall multimodal models. Even when coupled with lower-performing audio-based models, the text models lifted the combined model's performance to a more satisfactory level.

4.4. Qualitative Error Analysis

We conducted a qualitative error analysis to understand which types of sentence archetypes were frequently misclassified. We conducted this error analysis for (1) false positives vs. false negatives, (2) augmented vs. non-augmented sentences, and (3) short sentences removed vs. short sentences augmented.

False Positives: Our text model tended to misclassify certain types of sentences from control patients as originating from dementia patients. These sentences generally exhibited one or more of the following characteristics:

- **Noun-Phrase Sentences**: Examples include 'curtain on the window,' 'down on this side of the picture.'
- **Ungrammatical Sentences**: Sentence types that are uttered by patients in the control group but are slightly unnatural. Examples include 'the boy is uh taking cookies out of the cookie jar,' and 'uh mother's drying dishes,' 'that's real good then.'
- **Repetition**: The repetition of patients' sentences from the investigator, 'climbing a stool.'

False Negatives: Sentences from dementia patients that were misclassified as from control groups include:

- **Correct and Transcribed:**Sentences that were grammatically correct and transcribed correctly. Examples include 'that's about all,' and 'and the girl.'
- **Short and Correct:**Examples include sentences like 'here,' 'okay.'
- **Common responses:**Sentences that patients often respond to or ask and are transcribed correctly 'okay,' 'that's terrible,' 'that's about it, right?'

In the original-augmented model, the misclassified sentences are as follows:

- **Unlikely connotations:** Augmented sentences sometimes led to unlikely or misleading connotations, affecting the model's prediction accuracy.

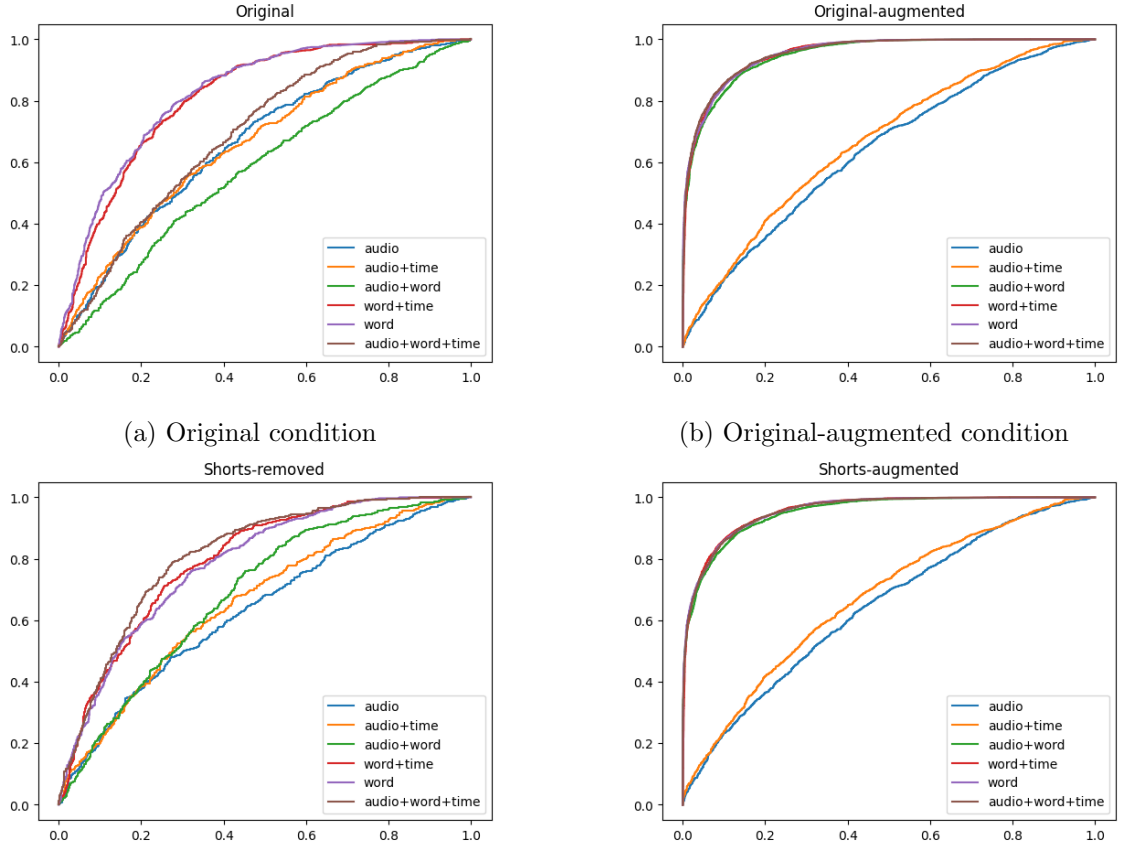
✓ 'I've got the tape recorder on so.'
(original, control, predicted as control)

✓ 'I've got the videotape recorder on so.'
(augmented, control, predicted as dementia)

✓ I've got the tape registrar on so.'
(augmented, control, predicted as dementia)

- **Word Usage:** Augmented words common to control data were sometimes present in sentences from dementia patients, leading to misclassification.

1. Note that the Audio+Text+Time model we saved had an exceptional performance, but the averaged performance was worse



(a) Original condition

(b) Original-augmented condition

(c) Shorts-removed condition

(d) Shorts-augmented condition

Figure 7: AUC-ROC curves in four conditions

✓ 'It shows the mother in the kitchen wiping dishes.'
(original, dementia, predicted as dementia)

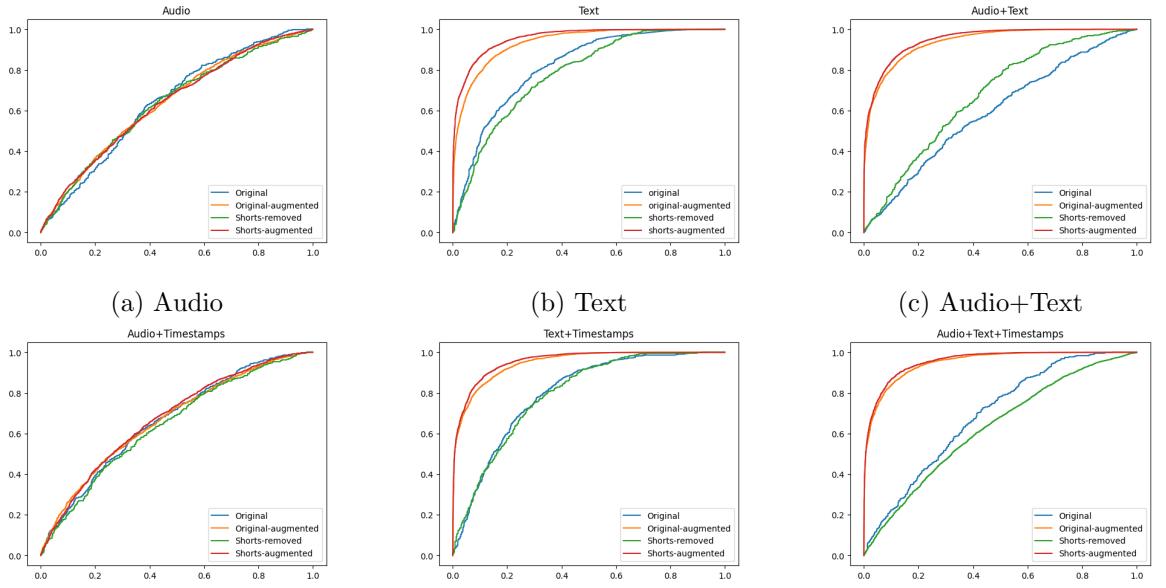
✓ 'It **testify** the mother in the kitchen wiping dishes.'
(augmented, dementia, predicted as control)

- **Augmented Correct Sentences:** Sentences that were originally correct but became incorrect after augmentation. For example:

✓ 'The little girl's standing there.'
(original, dementia, predicted as dementia)

✓ 'The little miss standing there.'
(augmented, dementia, predicted as control)

We also analyzed the model's performance under conditions where short sentences were either removed or augmented. The Shorts-removed condition revealed that the nature of incorrectly predicted sentences generally parallels that of the original dataset, minus the influence of short sentences. This suggests that the presence or absence of short sentences in the data does not dramatically alter the types of errors the model makes, which implies that the model's predictive capabilities are not significantly affected by sentence length alone. Interestingly, the errors made by the model in the Shorts-augmented condition closely resembled those in the original-augmented condition. This might point toward the robustness of the augmented dataset's influence on model behavior, regardless of the presence or absence of short sentences. This further emphasizes that while data augmentation significantly enhances the model's overall performance, it does not necessarily change the nature of the mistakes made by the model in prediction.



(a) Audio (b) Text (c) Audio+Text

(d) Audio+Timestamps (e) Text+Timestamps (f) Audio+Text+Timestamps

Figure 9: AUC-ROC curves of six types of models

5. Conclusion

In this study, we have presented a novel approach to dementia detection by leveraging multimodal data consisting of audio, text, and timestamps. Utilizing pre-trained models like Wav2vec and Word2vec, we have demonstrated robust performance when the model is sufficiently trained on a large dataset. Importantly, the presence of text data seems to bolster the performance of the model significantly, even compensating for lower-performing audio data. This suggests that text-based data can be a critical component in improving the diagnostic accuracy of dementia detection systems.

The study also highlights the limitations of individual modalities and the importance of integrating multiple types of data for more accurate results. While time-stamped data did not significantly impact the model's performance, the benefits of a multimodal approach are clear, especially when one data type alone might not be sufficient for accurate diagnosis.

Overall, this work sets the stage for further research into how multimodal data can be effectively used for medical diagnostics. The methodology we have developed could potentially provide a more effective and earlier diagnosis of dementia, offering a significant contribution to healthcare and quality of life.

Acknowledgments

We acknowledge the grants NIA AG03705 and AG05133 for supporting the development of the DementiaBank, Pitt Corpus. The technical support and advanced computing resources from University of Hawaii Information Technology Services – Cyberinfrastructure, funded in part by the National Science Foundation CC awards 2201428 and 2232862, are gratefully acknowledged.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: Comparing speech and language-based approaches

for alzheimer's disease detection. 07 2020. doi: 10.21437/Interspeech.2020-2557.

JT Becker, F Boller, OL Lopez, J Saxton, and KL McGonigle. The natural history of alzheimer's disease. description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6): 585—594, June 1994. ISSN 0003-9942. doi: 10.1001/archneur.1994.00540180063015. URL https://doi.org/10.1001/archneur.1994.00540180063015.

Karol Chlasta and Krzysztof Wołk. Towards computer-based automated screening of dementia through spontaneous speech. *Frontiers in Psychology*, 11, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.623237. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2020.623237.

Francois Chollet et al. Keras, 2015. URL https://github.com/fchollet/keras.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3: 642517, 04 2021. doi: 10.3389/fcomp.2021.642517.

Zhiqiang Guo, Zhaoci Liu, Zhen-Hua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. Text classification by contrastive learning and cross-lingual data augmentation for alzheimer's disease detection. pages 6161–6171, 01 2020. doi: 10.18653/v1/2020.coling-main.542.

Anna Hlédiková, Dominika Woszczyk, Alican Akman, Soteris Demetriou, and Björn Schuller. Data augmentation for dementia detection in spoken language, 2022.

Loukas Ilias, Dimitris Askounis, and John Psarras. A multimodal approach for dementia detection from spontaneous speech with tensor fusion layer. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, sep 2022. doi: 10.1109/bhi56158.2022.9926818. URL https://doi.org/10.1109%2Fbhi56158.2022.9926818.

Sweta Karlekar, Tong Niu, and Mohit Bansal. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2110. URL https://aclanthology.org/N18-2110.

Alyssa M. Lanzi, Anna K. Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L. Cohen. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438, 2023. doi: 10.1044/2022_AJSLP-22-00281. URL https://pubs.asha.org/doi/abs/10.1044/2022_AJSLP-22-00281.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge, 2020.

Lovro Matošević and Alan Jović. Accurate detection of dementia from speech transcripts using roberta model. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1478–1484, 2022. doi: 10.23919/MIPRO55190.2022.9803462.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity, 2020.

Iván G. Torre, Mónica Romero, and Aitor Álvarez. Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for english and spanish. *Applied Sciences*, 11(19), 2021. ISSN 2076-3417. doi: 10.3390/app11198872. URL https://www.mdpi.com/2076-3417/11/19/8872.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1670. URL https://aclanthology.org/D19-1670.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Ward Church. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech*, 2020. URL https://api.semanticscholar.org/CorpusID:226205766.