FUSING SPEECH AND LANGUAGE MODELS FOR DEMENTIA DETECTION

Tobias Deußer[*†,1,2], **Abdul Mohsin Siddiqi** [†,1], **Lorenz Sparrenberg** [1], **Tobias Adams** [1], **Christian Bauckhage** [1,2], **and Rafet Sifa** [1,2]
[1]University of Bonn, Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin, Germany
[†]*These authors contributed equally to this work.*

ABSTRACT

Accurate detection of dementia is crucial for timely intervention and care, and leveraging multimodal data holds significant potential for improving diagnostic accuracy. In this study, we explore deep learning approaches for dementia classification using the Pitt corpus, which includes brief participant descriptions of a cookie theft scene. We analyze 242 control and 307 dementia audio clips to investigate various representation learning techniques. Our best-performing approach fuses audio spectrograms with advanced language models, including Whisper model transcriptions and transformer-based feature extraction. We rigorously evaluate these models and find that our multimodal approach with an F_1-score of 86.42% eclipses other single modality approaches by a considerable margin. Our findings underscore the promise of multimodal deep learning techniques in advancing the reliability of dementia detection through audio analysis, possibly paving the way for more robust and accessible diagnostic tools.

Keywords Dementia Detection · Neurodegenerative Diseases · Multimodal Learning · Deep Learning · Machine Learning

1 Introduction

Dementia, particularly Alzheimer's disease (AD), represents a growing global health challenge. The number of dementia cases is projected to surge from 57.4 million in 2019 to 152.8 million by 2050, largely driven by the aging global population [1]. Dementia is not a single disease but a collection of symptoms marked by a decline in cognitive abilities relative to an individual's previous level of functioning [2]. It includes several distinct types, such as Alzheimer's disease (AD), Vascular Dementia (VaD), Lewy Body Dementia (LBD), Frontotemporal Dementia (FTD), and Mixed Dementia (MD) [3], with AD being the most prevalent, accounting for 60-70% of all dementia cases [4]. AD is marked by the buildup of intracellular neurofibrillary tangles and extracellular β-amyloid plaques, along with widespread synaptic loss and neuronal atrophy in the brain [5]. These neuropathological changes can begin years before clinical symptoms manifest [6]. However, a definitive diagnosis of AD can only be established through microscopic examination of brain tissue, typically during an autopsy. As a result, the term Dementia of the Alzheimer's Type (DAT) is used to refer to suspected cases of AD that have not yet been clinically confirmed [7]. Early and accurate diagnosis of DAT is crucial for improving patient outcomes, emphasizing the need for accessible diagnostic methods capable of detecting cognitive impairments in the earliest stages. Traditional diagnostic approaches—based on clinical assessments and neuropsychological testing—are often time-consuming and subject to variability in interpretation. This has driven growing demand for automated, scalable diagnostic solutions that offer greater consistency, efficiency, and earlier detection [8, 9, 10, 11, 12].

[*]tdeusser@uni-bonn.de, ORCID-ID: 0000-0003-4685-0847

While memory loss is often regarded as the primary symptom of DAT, language provides a valuable source of clinical information as well. Speech analysis, in particular, presents a promising non-invasive and cost-effective method for detecting cognitive decline. Linguistic and paralinguistic features of speech—such as fluency, articulation, and prosody—are well-established markers of dementia, offering insights into the early signs of cognitive impairment [13, 14, 15]. Recent advancements in machine learning and multimodal representation learning, which combine acoustic and linguistic features, have demonstrated significant potential for improving the accuracy of dementia detection.

In this paper, we utilize the Pitt dataset [16], which contains speech samples from both diagnosed dementia patients and a control group, to evaluate multimodal approaches. For transcription, we employ Whisper [17], while BERT [18] and Stella[1] are used for language representation. Additionally, spectrograms are utilized to extract audio features. Our goal is to improve dementia classification performance by comparing a variety of machine learning models, with a particular focus on integrating both audio and text features for a more comprehensive analysis.

Our multimodal dementia detection system improves upon existing single modality systems like [19] or [20] by a considerable margin. We achieve a Classification Accuracy of 86.59% and F_1-Score of 86.42% on our hold-out test set, demonstrating the feasibility of our approach.

In the following sections, we first discuss related studies. Section 3 describes our methodology. Thereafter, we highlight our experiments and results in Section 4. We close this paper with a conclusion and an outlook on future work.

2 Related Work

Several studies have explored the use of machine learning to automatically detect dementia and cognitive decline. Early approaches employed classical machine learning techniques, such as decision trees or Naive Bayes classifiers [21, 22, 23].

Modern dementia detection systems predominantly rely on deep neural networks and can be broadly categorized into three primary modalities: imaging data, clinical variables, and voice/language data [3].

Imaging data, particularly magnetic resonance imaging (MRI), has been extensively studied in dementia detection [24, 25, 26]. Image-based methods often outperform other modalities in terms of accuracy. For instance, one study reported 97% accuracy in multi-class Alzheimer's disease (AD) stage classification using the ADNI dataset and a pre-trained VGG19 model [27]. Despite their high performance, imaging-based techniques require expensive hardware and the involvement of skilled professionals for data acquisition and interpretation, which limits their practical application.

Clinical variables constitute the second major modality in dementia detection. These approaches typically involve cognitive assessments such as the Mini-Mental State Examination (MMSE) [28], as well as the analysis of biomarkers like amyloid, p-tau, and t-tau, often integrated with demographic factors such as age and gender [29]. Some studies also incorporate mRNA-based biosignatures [30] or leverage socio-demographic, basic health, and cognitive reserve proxy data [31]. While these methods have demonstrated promising results, they require specialized testing and expert interpretation, limiting their feasibility for routine, large-scale implementation.

The third modality focuses on voice and language data, which stands out due to its simplicity, non-invasive nature, and minimal hardware requirements. Initial efforts to utilize speech and language features for dementia detection were pioneered in studies like [32], [33], and [34], which laid the groundwork for future research in this area. More recent studies have shown that both linguistic and paralinguistic features of speech can be powerful indicators of dementia [35, 36, 37, 38, 39, 40, 41]. To establish standardized benchmarks for voice/language data, the ADReSS Challenge was introduced by [42], aiming to standardize Alzheimer's dementia detection through spontaneous speech analysis. This challenge provided a benchmark dataset and tasks for dementia classification and MMSE score regression, reinforcing the need for more robust models to improve upon existing techniques. Moreover, [19] demonstrated that purely acoustic features, particularly paralinguistic ones, could yield competitive accuracy in Alzheimer's detection using the Pitt [16] corpus. Their novel Active Data Representation (ADR) method achieved 78.70

Our work builds on these studies by integrating acoustic and textual features more effectively. We combine audio spectrograms with the Whisper transcription model [17] and use advanced language models to capture rich linguistic representations, aiming to improve dementia classification performance through a comprehensive multimodal approach.

For a more thorough review of recent advances in neurodegenerative disease detection using machine learning, we refer readers to the work of [43], [3], and [8].

[1] `https://huggingface.co/dunzhang/stella_en_1.5B_v5`

3 Methodology

We split this section into three parts: The first describes how we extract features from the audio signal, the second how we can leverage transcription models and language models to find a linguistic representation, and the third sheds light on how we combine these two methods to arrive at our final model architecture.

3.1 Audio

To convert the raw audio signal into a feature representation suitable for dementia detection, we compute the Mel-spectrogram. Given a one-dimensional audio signal with N_a total number of samples,

$$x[n], \quad n = 0, 1, 2, ..., N_a - 1, \tag{1}$$

we first divide it into overlapping frames. Each frame is windowed using a Hann window $w[m]$ [44], defined as:

$$x_w[m] = x[n + m] \cdot w[m], \tag{2}$$

where m is the index of the time sample within the current windowed frame, ranging from 0 to $L - 1$, and $w[m]$ is the Hann window function:

$$w[m] = 0.5 \left(1 - \cos\left(\frac{2\pi m}{L - 1}\right)\right) \tag{3}$$

and L is the window length.

For each windowed frame, we compute the Short-Time Fourier Transform (STFT) to obtain the frequency domain representation:

$$X_w[k] = \sum_{m=0}^{L-1} x_w[m] \cdot e^{-i2\pi \frac{km}{L}}, \quad k = 0, 1, \ldots, K - 1 \tag{4}$$

where k corresponds to the frequency bins. The magnitude spectrogram is obtained by taking the magnitude of the Fourier coefficients:

$$S_w[k] = |X_w[k]|, \tag{5}$$

where $S_w[k]$ represents the magnitude of the frequency component at bin k for the current windowed frame.

Next, we apply a set of M triangular filters spaced on the Mel scale [45] to transform the linear frequency f axis to the Mel scale, which is defined as:

$$M(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

Each Mel filter sums the energy in its corresponding frequency band:

$$S_m[j] = \sum_{k=0}^{K-1} S_w[k] \cdot H_j[k], \tag{7}$$

where $H_j[k]$ is the j-th Mel filter. The resulting Mel-spectrogram is a time-frequency representation where each row corresponds to a Mel frequency band and each column corresponds to a time frame. To compress the dynamic range of the values, we apply logarithmic scaling:

$$S_{\log}[j, t] = \log(S_m[j, t] + \epsilon), \tag{8}$$

where t is the index of the time frame and ϵ is a small constant to avoid taking the logarithm of zero. This Mel-spectrogram is then used as an input feature vector for our classification models.

3.2 Language

Similar to the previous section, we start with a raw audio signal. To compute a linguistic representation, we first transcribe the audio to text using a transcription model, namely Whisper [17]. This results in a string representation s, which is tokenized into sub-word units w_i, totaling N_l tokens. We then generate an embedding matrix E for the sentence s by passing the tokenized sequence s, represented as a vector of size N_l (with input IDs), through an "encoder-only" transformer model [46] $T(\cdot)$, which has an embedding size of T_e.

$$\mathrm{E} = T(s), \tag{9}$$

where E is the resulting matrix of size $N_l \times T_e$, with N_l corresponding to the number of tokens and T_e to the embedding size. This matrix is then the input to a pooling function $p(\cdot)$, such as *max* pooling, *mean* pooling, or *CLS* pooling, which aggregates the matrix along the first dimension. The resulting vector of dimension T_e is the second component of the feature vector used for our classification model.

3.3 Complete Model Architecture

We combine the models described in the previous sections by concatenating both representations and adding a classification head on top of these. We test a random forest and a multilayer perceptron as classifiers. The model architecture and training approach is shown in Figure 1.

4 Experiments

In this section, we outline the experiments conducted, the datasets utilized, and the results obtained. The model architecture employed is described in detail in Section 3. For hyperparameter optimization, we utilize Optuna [47], and for the classification task, we apply cross-entropy loss. All experiments were executed on a shared computing cluster equipped with four Nvidia A40 GPUs (48 GB VRAM each), two Intel Xeon 4309Y CPUs, and 400 GB of RAM.

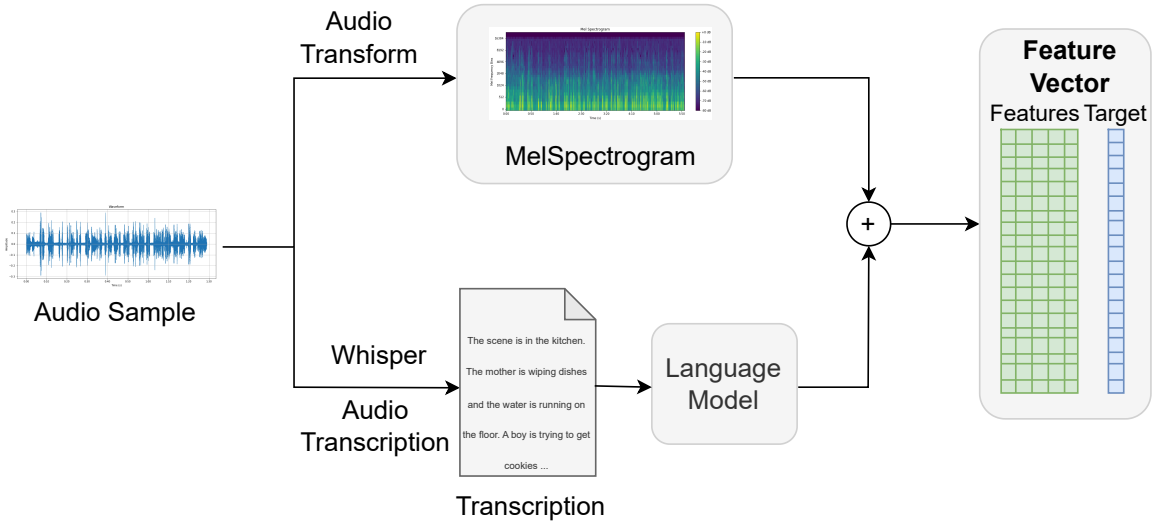
4.1 Data

The dataset used in this study is sourced from the Pitt corpus [16], a widely recognized resource from the University of Pittsburgh's Alzheimer's Research Program. It consists of audio recordings of spontaneous speech from participants categorized into two groups: Dementia and Control (healthy). The participants in the study are 44 years or older, have at least seven years of formal education, no history of nervous system disorders, and are not taking neuroleptic medications that could affect cognition. Additionally, all participants had an initial Mini-Mental State Examination (MMSE) score of 10 or higher [28], ensuring that they could provide reliable and meaningful speech data.

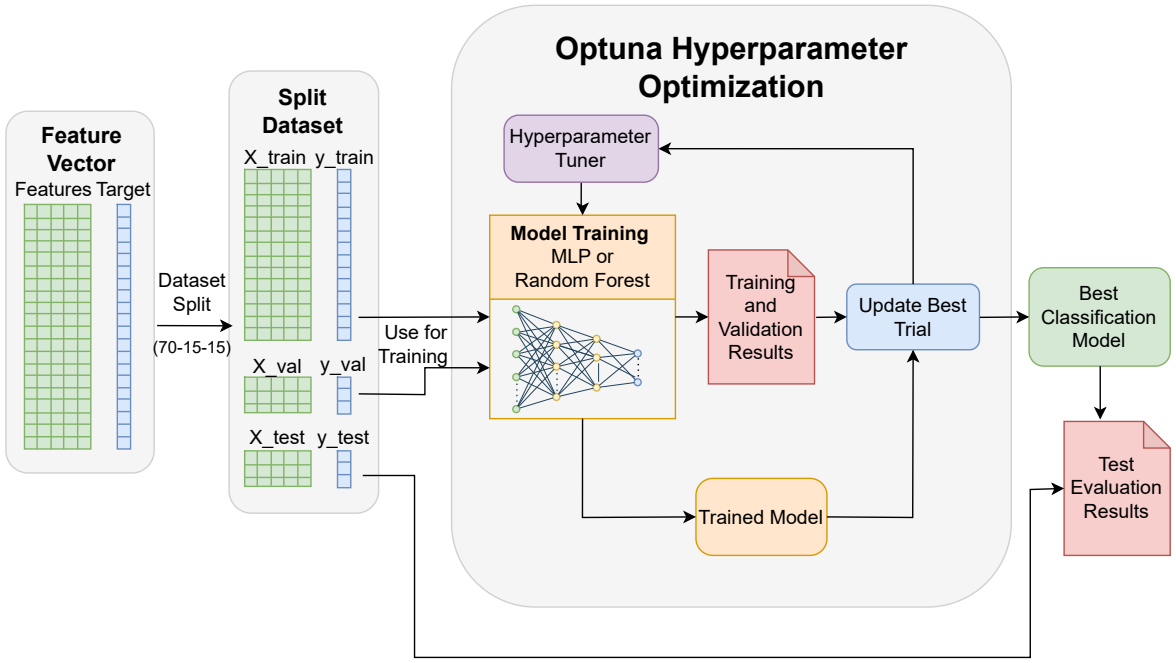
The dataset includes 99 control participants and 194 dementia patients, each contributing varying numbers of speech recordings across multiple visits. Control participants provided a total of 242 recordings, while dementia patients contributed 307. The demographic distribution of the dataset reflects a balanced representation of males and females, with most participants falling within the 65–70 and 70–75 age ranges. Both groups participated in a variety of tasks, including the Cookie Theft picture description [16], fluency exercises, story recall, and sentence construction. These tasks varied in duration, ranging from brief 1–2 minute activities to extended story recall sessions lasting up to an hour. The diversity of verbal tasks in this dataset enables a comprehensive analysis of cognitive and linguistic functions in both healthy individuals and those with dementia.

4.2 Results

To evaluate the various configurations and combinations outlined in Section 3 and depicted in Figure 1, we conducted a total of 42 experiments. The corresponding results are presented in Table 1. We began by testing a configuration that utilizes only the representation generated by the spectrogram, as described in Equations (1) to (8). Next, we evaluated five different Whisper models ranging from tiny to large [17] across two transformer architectures, Stella and BERT Base [18], yielding 10 configurations in total. For this, the transcriptions generated by the Whisper models are passed through the corresponding transformer model (See Equation (9)) to get the representations for classification. Finally, we combined the spectrogram-based and language model-based approaches into a multimodal system that integrates both audio and text features. This full model, as illustrated in Figure 1, combines the spectrogram with a transcribed text input processed by a language model. The results for this multimodal system are shown in the final two blocks of Table 1. We repeated the entire process using two different classifiers: a random forest (Table 1a) and a multilayer perceptron (Table 1b), allowing us to compare performance across these classifier architectures.



(a) Creating a feature vector from an audio sample.



(b) Using the created feature vector to train a model, optimize hyperparameters, and determine the best classification model.

Figure 1: Feature vector creation and model training and optimization. We use Optuna [47] as the framework for hyperparameter optimization.

As shown in Table 1, our system's performance improves significantly when multimodality is introduced by combining the Mel-Spectrogram with a language model using transcribed text. The best-performing configuration, which integrates

Stella, Whisper, and the Mel-Spectrogram, achieves an impressive F_1 score of 86.42% and an accuracy of 86.59%, substantially outperforming single-modality approaches such as [19], which reported a classification accuracy of 78.70%. Additionally, as expected, configurations using a Multilayer Perceptron proved to be considerably more powerful than those utilizing a Random Forest classifier.

5 Conclusion

This study introduced a multimodal deep learning approach for dementia detection using the Pitt [16] dataset, which is a collection of audio samples made by dementia patients and a control group. By combining audio spectrograms with language models, specifically, Whisper [17] for transcription and transformer-based models like BERT [18] and Stella for linguistic feature extraction, we developed a comprehensive framework that captures both acoustic and linguistic markers indicative of dementia.

Our experiments demonstrated that the multimodal approach significantly outperforms single-modality methods. The integration of Whisper transcriptions with a Stella model consistently yielded higher Classification Accuracy and F_1 scores compared to models utilizing only audio spectrograms or text features. The best-performing configuration achieved a classification accuracy of 86.59% and an F_1-Score of 86.42% on the hold-out test set, marking a substantial improvement over previous studies that relied solely on acoustic or linguistic features.

These results highlight the importance of capturing the multifaceted nature of dementia, affecting both speech patterns and the use of language. By effectively combining acoustic and linguistic features, our approach provides a more complete understanding of the patient's condition, which is crucial for early detection and intervention.

Future work could add another, crucial modality into our multimodal system: neuroimaging. A plethora of studies have found it effective in predicting dementia [48, 49, 50], which leads us to believe that adding neuroimages to our classification system would improve its predictive power even further. Additionally, evaluating the system's performance in real-world clinical settings will be essential to determine its practical applicability and scalability.

Furthermore, we plan to test various additional configurations, like adding a Tempogram [51] to include rhythmic analysis or augmenting the audio data, e.g., time stretching, pitch shifting, and adding background noise, to improve the stability of our models and size of the dataset. In a similar vein, we plan to utilize and test our system on other corpora available at the DementiaBank [52], possibly in other languages, like the Spanish or Mandarin one introduced in [53] and [54], respectively.

In conclusion, this research demonstrates the efficacy of multimodal deep learning techniques in enhancing the reliability of dementia detection through audio and linguistic analysis. The proposed approach holds promise for developing more robust and accessible diagnostic tools, contributing to earlier interventions and improved patient outcomes.

Acknowledgments

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

References

[1] Emma Nichols, Jaimie D Steinmetz, Stein Emil Vollset, Kai Fukutaki, Julian Chalek, Foad Abd-Allah, Amir Abdoli, Ahmed Abualhasan, Eman Abu-Gharbieh, Tayyaba Tayyaba Akram, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. *The Lancet Public Health*, 7(2):e105–e125, 2022.

[2] Alexander Kurz, Hans-Jürgen Freter, Susanna Saxl, and Ellen Nickel. *Demenz. Das Wichtigste*. Deutsche Alzheimer Gesellschaft e. V., Berlin, 8th edition, 2019.

[3] Ashir Javeed, Ana Luiza Dallora, Johan Sanmartin Berglund, Arif Ali, Liaqata Ali, and Peter Anderberg. Machine learning for dementia prediction: a systematic review and future research directions. *Journal of medical systems*, 47(1):17, 2023.

[4] World Health Organization. Dementia fact sheet. `https://www.who.int/news-room/fact-sheets/detail/dementia`, March 2023. Accessed: 2024-10-15.

[5] Dennis J Selkoe and John Hardy. The amyloid hypothesis of alzheimer's disease at 25 years. *EMBO Molecular Medicine*, 8(6):595–608, 2016. doi:https://doi.org/10.15252/emmm.201606210.

[6] Elizabeth W Twamley, Susan A Legendre Ropacki, and Mark W Bondi. Neuropsychological and neuroimaging changes in preclinical alzheimer's disease. *Journal of the International Neuropsychological Society*, 12(5): 707–735, 2006. doi:10.1017/S1355617706060863.

[7] Robert E Hales, Stuart C Yudofsky, and Glen O Gabbard. *American Psychiatric Publishing Textbook of Psychiatry*. American Psychiatric Publishing Inc, Arlington, VA, 5th edition, 2008. ISBN 9781585622573.

[8] Nikhil Pateria and Dilip Kumar. A comprehensive review on detection and classification of dementia using neuroimaging and machine learning. *Multimedia Tools and Applications*, 83(17):52365–52403, 2024.

[9] Elena Doering, Merle C. Hönig, Tobias Deußer, Gérard N. Bischof, Thilo van Eimeren, Alexander Drzezga, and Lotta M. Ellingsen. Translating the future: image-to-image translation for the prediction of future brain metabolism. In *Medical Imaging 2024: Clinical and Biomedical Imaging*. International Society for Optics and Photonics, 2024.

[10] Pranav Mahajan and Veeky Baths. Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience*, 13:623607, 2021.

[11] Suriya Murugan, Chandran Venkatesan, MG Sumithra, Xiao-Zhi Gao, B Elakkiya, Muthuramalingam Akila, and S Manoharan. Demnet: A deep learning model for early diagnosis of alzheimer diseases and dementia from mr images. *IEEE Access*, 9:90319–90329, 2021.

[12] Ploypaphat Saltz, Shih Yin Lin, Sunny Chieh Cheng, and Dong Si. Dementia detection using transformer-based deep learning and natural language processing models. In *Proc. ICHI*, pages 509–510. IEEE, 2021.

[13] Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195, 2015.

[14] Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015:1–15, 2015.

[15] Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113, 2021.

[16] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594, 1994. Supported by Grants NIA AG03705 and AG05133.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proc. NAACL-HLT*, pages 4171–4186, 2019. doi:10.18653/v1/N19-1423.

[19] Fasih Haider, Sofia de la Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, pages 272–281, 2020.

[20] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422, 2015.

[21] Piew Datta, WR Shankle, and Michael Pazzani. Applying machine learning to an alzheimer's database. In *Proc. AAAI symposium*, pages 25–27, 1996.

[22] William Rodman Shankle, Subramani Mani, Michael J Pazzani, and Padhraic Smyth. Detecting very early stages of dementia from normal aging with machine learning methods. In *Proc. AIME*, pages 71–85. Springer, 1997.

[23] Subramani Mani, Malcolm B Dick, Michael J Pazzani, Evelyn L Teng, Daniel Kempler, and I Maribell Taussig. Refinement of neuro-psychological tests for dementia screening in a cross cultural population using machine learning. In *Proc. AIMDM*, pages 326–335. Springer, 1999.

[24] EL-Geneedy Marwa, Hossam El-Din Moustafa, Fahmi Khalifa, Hatem Khater, and Eman AbdElhalim. An mri-based deep learning approach for accurate detection of alzheimer's disease. *Alexandria Engineering Journal*, 63:211–221, 2023.

[25] Pierluigi Carcagnì, Marco Leo, Marco Del Coco, Cosimo Distante, and Andrea De Salve. Convolution neural networks and self-attention learners for alzheimer dementia diagnosis from brain mri. *Sensors*, 23(3):1694, 2023.

[26] Sven Haller, Hans Rolf Jäger, Meike W Vernooij, and Frederik Barkhof. Neuroimaging in dementia: more than typical alzheimer disease. *Radiology*, 308(3):e230173, 2023.

[27] Hadeer A Helaly, Mahmoud Badawy, and Amira Y Haikal. Deep learning approach for early detection of alzheimer's disease. *Cognitive Computation*, 14:1711—1727, 2022. doi:10.1007/s12559-021-09946-2.

[28] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

[29] Daniel Stamate, Min Kim, Petroula Proitsi, Sarah Westwood, Alison Baird, Alejo Nevado-Holgado, Abdul Hye, Isabelle Bos, Stephanie J.B. Vos, Rik Vandenberghe, et al. A metabolite-based machine learning approach to diagnose alzheimer-type dementia in blood: Results from the european medical information framework for alzheimer disease biomarker discovery cohort. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5(1):933–938, 2019. doi:https://doi.org/10.1016/j.trci.2019.11.001.

[30] Makrina Karaglani, Krystallia Gourlia, Ioannis Tsamardinos, and Ekaterini Chatzaki. Accurate blood-based diagnostic biosignatures for alzheimer's disease via automated machine learning. *Journal of Clinical Medicine*, 9 (9):3016, 2020. doi:10.3390/jcm9093016.

[31] David Facal, Sonia Valladares-Rodriguez, Cristina Lojo-Seoane, Arturo X. Pereiro, Luis Anido-Rifon, and Onésimo Juncos-Rabadán. Machine learning approaches to studying the role of cognitive reserve in conversion from mild cognitive impairment to dementia. *International Journal of Geriatric Psychiatry*, 34(7):941–949, 2019. doi:10.1002/gps.5090.

[32] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *Proc. ICMA*, volume 3, 2005.

[33] Bart Peintner, William Jarrold, Dimitra Vergyri, Colleen Richey, Maria Luisa Gorno Tempini, and Jennifer Ogar. Learning diagnostic models using speech and language measures. In *Proc. EMBC*, pages 4648–4651. IEEE, 2008.

[34] Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090, 2011.

[35] M. Rupesh Kumar, Susmitha Vekkot, S. Lalitha, Deepa Gupta, Varasiddhi Jayasuryaa Govindraj, Kamran Shaukat, Yousef Ajami Alotaibi, and Mohammed Zakariah. Dementia detection from speech using machine learning and deep learning architectures. *Sensors*, 22(23), 2022.

[36] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Markus Reuber, and Heidi Christensen. Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79, 2019.

[37] Ayimnisagul Ablimit, Catarina Botelho, Alberto Abad, Tanja Schultz, and Isabel Trancoso. Exploring dementia detection from speech: Cross corpus analysis. In *Proc. ICASSP*, pages 6472–6476. IEEE, 2022.

[38] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Speechformer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:775–788, 2023.

[39] Samad Amini, Boran Hao, Lifu Zhang, Mengting Song, Aman Gupta, Cody Karjadi, Vijaya B Kolachalama, Rhoda Au, and Ioannis Ch Paschalidis. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimer's & Dementia*, 19(3):946–955, 2023.

[40] Rui He, Kayla Chapin, Jalal Al-Tamimi, Núria Bel, Marta Marquié, Maitee Rosende-Roca, Vanesa Pytel, Juan Pablo Tartari, Montse Alegret, Angela Sanabria, et al. Automated classification of cognitive decline and probable alzheimer's dementia across multiple speech and language domains. *American Journal of Speech-Language Pathology*, 32(5):2075–2086, 2023.

[41] Laura C Maclagan, Mohamed Abdalla, Daniel A Harris, Therese A Stukel, Branson Chen, Elisa Candido, Richard H Swartz, Andrea Iaboni, R Liisa Jaakkimainen, and Susan E Bronskill. Can patients with dementia be identified in primary care electronic medical records using natural language processing? *Journal of Healthcare Informatics Research*, 7(1):42–58, 2023.

[42] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3, 2021.

[43] Ngumimi Karen Iyortsuun, Soo-Hyung Kim, Min Jhon, Hyung-Jeong Yang, and Sudarshan Pant. A review of machine learning and deep learning approaches on mental health diagnosis. In *Healthcare*, volume 11, page 285. MDPI, 2023.

[44] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[45] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proc. NeurIPS*, volume 30, 2017.

[47] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proc. KDD*, 2019.

[48] Hongyoon Choi, Kyong Hwan Jin, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural brain research*, 344:103–109, 2018.

[49] Abdul Rehman, Myung-Kyu Yi, Abdul Majeed, and Seong Oun Hwang. Early diagnosis of alzheimer's disease using 18f-fdg pet with soften latent representation. *IEEE Access*, 2024.

[50] E Doering, T Deußer, M Hoenig, T van Eimeren, L Ellingsen, and A Drzezga. How will you age? A glimpse into future brain aging on FDG-PET using deep learning. *Nuklearmedizin-NuclearMedicine*, 62(02):L7, 2023.

[51] Peter Grosche, Meinard Müller, and Frank Kurth. Cyclic tempogram—a mid-level tempo representation for musicsignals. In *Proc. ICASSP*, pages 5522–5525. IEEE, 2010.

[52] 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.

[53] Olga Ivanova, Juan José G Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E Llorente, and Nuria Carcavilla González. Discriminating speech traits of alzheimer's disease assessed through a corpus of reading task for spanish language. *Computer Speech & Language*, 73:101341, 2022.

[54] Guanyu Zhang, Jinghong Ma, Piu Chan, and Zheng Ye. Graph theoretical analysis of semantic fluency in patients with parkinson's disease. *Behavioural Neurology*, 2022(1):6935263, 2022.

Table 1: Results with various configurations.

(a) Results in % with a **Random Forest** as classifier.

Configuration	Accuracy	Precision	Recall	F_1-Score	AUROC
Mel-Spectrogram	54.88	55.54	57.77	56.63	46.43
BERTBase + WhisperTiny	68.29	68.81	72.70	70.70	76.67
BERTBase + WhisperBase	78.05	78.39	80.58	79.47	77.86
BERTBase + WhisperSmall	69.51	70.06	74.83	72.37	76.67
BERTBase + WhisperMedium	75.61	75.95	77.95	76.75	80.66
BERTBase + WhisperLarge	74.39	74.76	77.07	75.74	79.82
Stella + WhisperTiny	76.83	77.14	78.84	77.48	85.00
Stella + WhisperBase	78.05	78.33	79.75	79.04	89.29
Stella + WhisperSmall	75.61	75.95	77.95	76.75	86.55
Stella + WhisperMedium	76.83	77.14	78.84	77.48	85.95
Stella + WhisperLarge	76.83	77.14	78.84	77.48	87.14
Mel-Spectrogram + BERTBase + WhisperTiny	69.51	72.56	69.94	71.23	72.92
Mel-Spectrogram + BERTBase + WhisperBase	78.05	80.58	78.39	79.47	78.45
Mel-Spectrogram + BERTBase + WhisperSmall	70.73	75.65	71.25	73.38	77.98
Mel-Spectrogram + BERTBase + WhisperMedium	71.95	74.40	72.32	73.34	78.63
Mel-Spectrogram + BERTBase + WhisperLarge	73.17	76.19	73.57	74.86	78.45
Mel-Spectrogram + Stella + WhisperTiny	76.83	79.73	77.20	78.45	84.82
Mel-Spectrogram + Stella + WhisperBase	73.17	76.19	73.57	74.86	89.29
Mel-Spectrogram + Stella + WhisperSmall	75.61	77.19	75.89	76.53	86.31
Mel-Spectrogram + Stella + WhisperMedium	73.17	74.08	73.39	73.73	87.32
Mel-Spectrogram + Stella + WhisperLarge	79.27	80.67	79.52	80.09	88.93

(b) Results in % with a **Multilayer Perceptron** as classifier.

Configuration	Accuracy	Precision	Recall	F_1-Score	AUROC
Mel-Spectrogram	60.98	60.30	65.02	62.57	64.35
BERTBase + WhisperTiny	67.07	67.38	68.47	67.92	74.35
BERTBase + WhisperBase	76.83	77.08	78.14	77.61	76.67
BERTBase + WhisperSmall	73.17	73.45	74.63	74.04	77.68
BERTBase + WhisperMedium	76.83	76.73	76.97	76.85	82.98
BERTBase + WhisperLarge	78.05	78.16	78.29	78.22	80.00
Stella + WhisperTiny	81.71	81.91	82.58	82.24	86.37
Stella + WhisperBase	81.71	81.73	81.71	81.72	89.94
Stella + WhisperSmall	80.49	80.66	81.10	80.88	86.85
Stella + WhisperMedium	81.71	81.85	82.13	81.99	88.93
Stella + WhisperLarge	82.93	82.98	82.98	82.98	89.23
Mel-Spectrogram + BERTBase + WhisperTiny	73.17	73.21	73.21	73.21	79.64
Mel-Spectrogram + BERTBase + WhisperBase	81.71	81.71	81.73	81.72	79.94
Mel-Spectrogram + BERTBase + WhisperSmall	79.27	83.60	79.70	81.61	84.76
Mel-Spectrogram + BERTBase + WhisperMedium	81.71	81.71	81.73	81.72	83.60
Mel-Spectrogram + BERTBase + WhisperLarge	82.93	82.98	82.98	82.98	83.81
Mel-Spectrogram + Stella + WhisperTiny	81.71	82.13	81.85	81.99	85.83
Mel-Spectrogram + Stella + WhisperBase	86.59	87.79	86.37	86.42	87.98
Mel-Spectrogram + Stella + WhisperSmall	82.93	82.98	82.98	82.98	88.33
Mel-Spectrogram + Stella + WhisperMedium	84.15	84.60	84.29	84.13	89.82
Mel-Spectrogram + Stella + WhisperLarge	85.37	85.42	85.42	85.37	89.35