An AI-Assisted Approach for the Detection of Dementia

Progress Report

Stefania Livori
Department of Artificial Intelligence
University of Malta
Msida, Malta
stefania.livori.23@um.edu.mt

***Abstract*—This project aims to develop a multi-modal AI system for early dementia detection, focusing on speech and language analysis. The system will compare several AI frameworks and machine learning libraries to identify the most accurate and efficient model for dementia detection. These models will be taken from two specific research papers. The best-performing models will then be used to build a better system, aiming to achieve better performance, meaning a better structure in terms of layers, fewer computational resources, or higher accuracy, precision, and recall, and to investigate ethical concerns when choosing a single dataset. The final system will be a complete CNN architecture that produces a prediction used to determine whether there is dementia. The importance of this study is the detection of dementia and the need for a system to help psychologists and doctors detect dementia faster than before.**

I. Introduction

A. Introduction

The word "dementia" literally translates to "a mind without a mind" [1]. By 1822, the term became closely associated with "senile dementia," referring to age-related mental decline. In the late 19th century, "dementia praecox" referred to what is now called schizophrenia [1].

B. Problem Definition

Dementia is a serious and progressive neurological cognitive disorder that leads to decreased daily functioning, affecting millions of people around the world. Early detection is crucial to improving the management of symptoms and quality of life for both patients and caregivers. Having a timely diagnosis helps individuals and their families plan better and receive early interventions. But dementia is often missed in the early stages because mild symptoms resemble normal aging. This delay in recognition can make it much harder to manage daily activities and responsibilities [1].

C. Types of Dementia

The most common form, Alzheimer's disease, accounts for approximately 60–70% of cases and is characterized by the accumulation of amyloid plaques and tau tangles, leading to neuronal loss and gradual cognitive deterioration, with symptoms such as memory loss, confusion, disorientation, and language difficulties [1], [2]. Vascular dementia, the second most prevalent type, results from reduced blood flow to the brain, often following strokes or chronic small vessel disease, and is associated with executive dysfunction, impaired judgment, and movement-related cognitive decline [1], [3]. Dementia with Lewy bodies (DLB) affects around 10–15% of patients and is distinguished by abnormal alpha-synuclein deposits, causing fluctuating cognition, visual hallucinations, Parkinson's symptoms, and sleep disturbances [2], [4]. Frontotemporal dementia (FTD) primarily affects individuals around age 65, leading to frontal and temporal lobe degeneration, personality changes, inappropriate social behavior, impaired judgment, and language difficulties [4], [5]. Mixed dementia, commonly a combination of Alzheimer's and vascular dementia, is increasingly recognized and poses diagnostic challenges due to overlapping pathologies [3], [6].

II. Motivation

The project is propelled by the escalating prevalence of dementia, both globally and within Malta, where demographic projections indicate that the number of individuals affected will nearly double by 2050. Early screening is widely recognized to improve the quality of life for patients and their caregivers by facilitating timely intervention, planning, and support. Harnessing AI technologies presents a promising opportunity to deliver scalable, cost-effective, and accurate detection tools, potentially bridging gaps in access to early diagnosis and personalized care. AI-driven solutions can expand capacity for screening and prediction, alleviate pressure on health services, and empower patients and families to better manage dementia's complex challenges [7]–[10].

III. Background Research and Literature Review

A. Introduction

Dementia represents a major global health challenge. According to the World Health Organization (WHO), 60% to 70% of the world's population diagnosed with dementia amounts to approximately 47.5 million people [11], [12]. This number is projected to increase to 139 million by 2050 [7]–[9], [13]. Dementia is also a growing concern in the Maltese Islands. Two national strategies, between 2015 and 2031 [14] [15], have been implemented following interviews with individuals living with or caring for persons with dementia

[14]. Maltese statistics indicate that the number of people with dementia will double from 6,552 in 2018 to 14,117 in 2050 [16], highlighting the urgent need for effective interventions.

B. Action Areas in Dementia Strategy

The 2024–2031 Maltese Dementia Strategy outlines key action areas, including:

- Action Area 2: Decreasing the risk of dementia
- Action Area 3: Timely diagnosis

Early detection is emphasized as a critical factor for slowing disease progression [17]. Emphasizing the need for a study in dementia at the University of Malta, is research done by C. Cachia [18], which shows the need to detect whether people with dementia are at risk of hurting themselves.

C. Artificial Intelligence for Dementia Detection

1) Introduction: AI has shown considerable potential in addressing medical challenges. AI has also been investigated in the field of dementia detection. Early detection systems can aid elderly care and assist caregivers in planning for healthcare, legal, and financial decisions [1].

2) Feature Engineering for Alzheimer's Detection: Studies have applied feature engineering to classify speech as Alzheimer's Disease (AD) or non-AD [19]. Key indicators include hesitation words, repeated words, linguistic markers, and synonym replacement, which have shown strong correlations with cognitive decline [19]. Considering the technology and power that we have in 2025, this system can be made better by not assuming that everyone who has dementia suffers from saying hesitation or repeated words, but concatenate it with the frequency and vocal sounds using Multi-modal Learning Analytics (MMLA), which might not give better accuracy, but would be more powerful in terms of real-life performance.

D. MMLA in Alzheimer Detection

MMLA approaches combining audio, video, and physiological measures have shown increased accuracy in cognitive impairment detection. Current literature identifies challenges like limited standardization across studies. Most existing solutions remain restricted in real-world clinical settings.

IV. AIMS AND OBJECTIVES

In this Final Year Project, an AI-based system will be developed for the early detection of Alzheimer's disease using speech and language analysis. The project involves designing and evaluating a multi-modal model, integrating both audio and textual features. Due to the lack of publicly available datasets containing both modalities, audio recordings will be converted into spectrograms, while transcripts will be processed as text input. Several existing architectures from recent research will be implemented and compared to determine the most accurate and computationally efficient approach. The main deliverable will be a functional model capable of predicting early signs of Alzheimer's symptoms from speech, with potential future use as part of a clinical or assistive application.

The main objectives are:

- Evaluating multiple multi-modal AI models for detecting dementia-related speech patterns through experiments.
- Choosing the 3 best-performing detection models, and the best dataset.
- Build a better model architecture, using the same structure and architectural findings from these research papers.
- Noting down the findings of what makes a model better, which can then be used by a User Interface designer to build an early dementia detection application.

The insights gained from this project guide the design of AI-driven early screening tools for dementia, highlighting which multi-modal approaches are most effective and how interactive speech tasks can be applied in real-world settings.

V. PROPOSED SOLUTION AND METHODOLOGY

A. Datasets

1) Introduction: The datasets used for this experiment are two datasets, which are both freely available on the DementiaBank website upon access [20], the Cookie dataset from the Pitt Corpus [21], and the ADReSSo dataset [22]. In this part of the project, to replicate the results, I made sure that I used the same datasets as the original research.

2) The Pitt Corpus: The Pitt Corpus is one of the most widely used datasets for research on speech-based detection of AD [10]. It forms part of the DementiaBank database, developed under the TalkBank project to facilitate the study of language and communication disorders [18]. It includes speech samples from individuals diagnosed with Alzheimer's disease as well as healthy control participants, accompanied by demographic and clinical information such as Mini-Mental State Examination (MMSE) scores [10]. The Pitt Corpus serves as a valuable resource for developing and evaluating computational models that analyze linguistic and acoustic markers of cognitive decline. The manual transcriptions capture rich linguistic detail, including pauses, repetitions, and incomplete utterances, enabling fine-grained analysis of speech patterns [10]. Numerous studies have used this dataset to extract acoustic, prosodic, and linguistic features for automatic AD detection and progression monitoring [22]–[24]. Due to its well-annotated structure and public availability, the Pitt Corpus has become a benchmark dataset in the field, supporting reproducibility and cross-study comparison in the development of speech-based cognitive assessment systems [10], [23], [24].

3) The ADReSSO Dataset: The ADReSSo Challenge, introduced at INTERSPEECH 2020, provides a platform for researchers to develop and benchmark models for detecting cognitive decline using speech [22]. The challenge aims to systematically compare different approaches to detecting cognitive impairment from spontaneous speech [22] through three main tasks: AD Classification, MMSE Score Regression or Cognitive Decline Prediction [22]. To ensure fair comparison across methods, the challenge provides standardized datasets of spontaneous speech, which is readily split into testing and training sets [22]. A significant aspect of the ADReSSo

Challenge is its focus on analyzing raw audio directly, without relying on manual transcriptions, though automatic transcription is permitted for extracting linguistic features [22].

B. Objective 1: Evaluating two similar research papers using multi-modal techniques

1) Introduction: Two distinct models were designed to approach dementia detection models from complementary perspectives, one built on multi-modal fusion of text, timestamps, and audio using the Pitt Corpus, and another leveraging deep transformer-based multi-modal fusion through the ADReSSo dataset. Both licenses where freely available and open source upon referencing the model. The authors in fact encourage people to work upon their solutions and adapt the solution to gain better results [12], [25].

2) First Experiment: The first model, based on K. Lin and P. Y. Washington [12], focuses on integrating acoustic, linguistic, and temporal cues. It uses Wav2Vec2 [26], [27] for extracting robust speech features and Word2Vec2 [28] for textual embeddings, later combined through Long Short-Term Memory (LSTM) [29] networks that capture temporal dependencies. Timestamp alignment allows the model to incorporate response time and pacing as features, improving sensitivity to cognitive patterns. The approximate timestamps had to be used, since although the author stated that there should be timestamps, I did not manage to extract them. The architecture uses late fusion, followed by a sigmoid function via concatenation to combine features. The reason why this model is towards my interest is the fact that it uses timestamps which are readily available in the case of real-life speech, and is interesting to experiment on.

3) Second Experiment: The second model, from D. Altinok [25], the BertImage architecture, combines speech-derived Mel-spectrograms processed via Vision Transformer (ViT) [30] with Robustly Optimized BERT Pretraining Approach (RoBERTa) transcript embeddings for the same sample's transcript. These embeddings are fused using a cross-attention mechanism, allowing each modality to inform and enrich the other. The model performs two simultaneous tasks: binary classification for dementia detection and classification of MMSE score groups, providing more nuanced clinical insight. Training uses the AdamW optimizer, multi-task loss, and fine-tuning of top transformer layers for domain adaptation. Spectrogram generation and Whisper-based transcription pipelines ensure consistent input formatting.

4) Med-BERT: Med-BERT, developed by Rasmy et al. [31], is a domain-specific adaptation of the BERT model for medical data. Trained on a large diverse dataset of about 28 million patient records from the Cerner Health Facts database [31]. It captures semantic and temporal dependencies of medical codes in Electronic Health Records (EHRs). Med-BERT has demonstrated strong generalization across disease prediction tasks, including heart failure among diabetic patients. The model introduces interpretability tools that visualize attention mechanisms, enhancing transparency in clinical decision-making [31].

Med-BERT is trained on a large, diverse corpus of electronic health records (EHRs), incorporating a richer clinical vocabulary and modeling the temporal order of medical codes within patient visits. Its design enables robust cross-institutional deployment. Experimental evaluations demonstrate substantial improvements in predictive performance, especially in low-data settings, highlighting its effectiveness when labeled EHR data are scarce. Additionally, the study introduces interpretability tools that visualize attention mechanisms, promoting transparency and trust in clinical decision support systems.

C. Objective 2: Identify the three best-performing model and the dataset which must be used to train the model

At this point, I saw the results that were gained from each run and built a table for each different set, one for experiment 1 with the original dataset, one for experiment 1 with the augmented dataset, and one for experiment 2, and saw what accuracies and other metrics were achieved. The objective of this phase was mainly to evaluate the literature review that already exists. In the field of AI, it does not always make sense to build a neural network from the beginning, especially when there are build models already existing and the case studied is not an easily replicated case. For the first model, proposed by K. Lin and P. Y. Washington, multiple combinations of modalities were tested. The text-only model achieved the best results in the original dataset, achieving an accuracy of 62.28% and an AUC of 62.12%, but the audio, text, and time model performed best when using data augmentation via synonym replacement was applied, achieving an accuracy of 71.18% and an AUC of 77.7%. However, this model exhibited some limitations. The preprocessing relied heavily on approximated word-level timestamps, and the synonym replacement technique sometimes introduced unrealistic phrases. Additionally, potential data leakage across speaker samples might have inflated accuracy. Despite these issues, the model provided a strong foundation for multi-modal learning and highlighted the effectiveness of combining linguistic and acoustic information. The second experiment, which is the BertImage model, based on D. Altinok's research [25], used the ADReSSo dataset and introduced a more advanced multi-modal fusion mechanism using ViT [30] for Mel-spectrograms and RoBERTa [32] for transcript embeddings. Cross-attention layers were used to integrate visual and textual modalities, and the system simultaneously predicted binary dementia classification and MMSE cognitive score categories. The cross-attention fusion architecture allowed each modality to attend to the other, enriching both linguistic and acoustic feature representations. The model achieved an average precision of 0.85 and a recall of 0.83 for binary dementia detection, consistent with results from the original ADReSSo challenge [22]. On the other hand, the MMSE prediction and label plus MMSE prediction, which were not provided but added for experimentation, performed very badly, especially for class 1, giving a precision, accuracy, recall, and F1-Score of 0%, which is totally undesirable, regardless of the good performance in the other classes.

D. Objective 3: Building a model for Dementia Classification

At an earlier stage, I wanted to build an application upon the best-performing model from the evaluations made. When having an introductory informal meeting with the dementia intervention team (DIT), it was discussed that for a person to be diagnosed with dementia, it takes at least three psychologists to examine the patients. Therefore, as a result, I plan to examine further exactly how the architecture is built and address the needs in this field of study.

This phase started by running the whole model from the second experiment by D. Altinok [25] using the Pitt Corpus [21]. Figure 1 and 2 shows, the data flow of the original model.

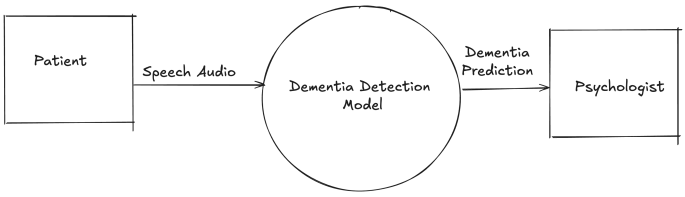

Fig. 1. DFD Level 0.

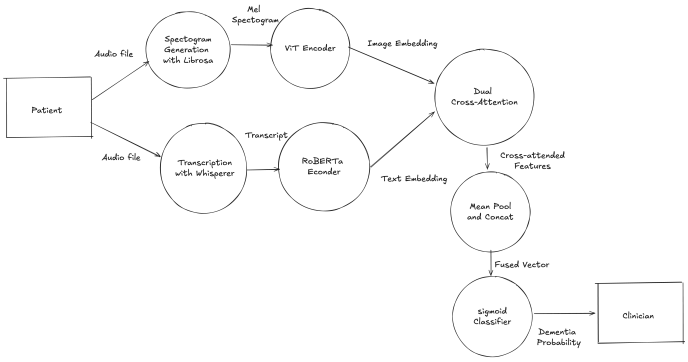

Fig. 2. DFD Level 1.

I first needed to clear the code from MMSE variables since they are not available in the Pitt Corpus [21], and did not give a good result in the second experiments. This was quite an easy step where any unnecessary computation which were totally not used by the model were removed.

Secondly, I needed to preprocess the dataset by turning the audio files into spectrograms and transcripts, using the code that was used in the literature review by the second experiment. This was done with the same code that was provided and emailed to me by D. Altinok [25] upon explanation about how the spectrogram and transcripts where built.

Unlike the ADReSSo dataset, the Pitt Corpus is not readily available split into a training and testing set, therefore the whole dataset was split into a 20:80 ratio testing and training respectively, where the training was once again split into a 20:80 ratio of validation and training set respectively, this split was the one used by K. Lin and P. Y. Washington [12].

As an enhancement to the computational time and performance, I precomputed the results for the spectrogram transformation and text tokenization, which was done for every epoch. This means that beforehand, for each epoch, the spectrograms and text were preprocessed. This was done to minimize the computational time needed for the 30 epochs to run.

Once all these steps were done, I have started experimenting, adding or removing some layers. Parts from the other two models, which performed the best in the first experiment will also be experimented by adding them into the architecture. Therefore, the main architecture will now be the best performing model from D. Altinok's research [25], and experimenting, adding to it other layers from the two models, which performed best in the first experiment, which do not have data augmentation.

E. Objective 4: Noting down the findings and evaluating the final model results

In this stage, I will be noting down any improvements that were made to the model and how they were made. This does not just include accuracy, but also performance, computational time, and architectural design.

I will also design a new architecture diagram to better illustrate my model, based on the original author's design.

VI. EVALUATION

A. Objective 1: Evaluating two similar research papers using multi-modal techniques

1) Introduction: Each model was evaluated under a rigorous, multi-level framework designed to ensure validity, fairness, and performance interpretability.

2) First Experiment: For the Pitt Corpus model, a 5-fold stratified cross-validation was performed to maintain balanced dementia-control proportions. The evaluation metrics included accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). Validation loss directed early stopping to prevent overfitting. Original and augmented datasets were compared parametrically to assess performance improvements from synonym-based text augmentation [12].

3) Second Experiment: The ADReSSo BertImage model underwent classification using both binary classification for label prediction only and cross-entropy classification evaluations using multi-task training. Precision, recall, and macro/weighted F1-scores were computed separately for each dementia label classification and MMSE category prediction. Confusion matrices highlighted class imbalance effects, particularly lower performance in MMSE bins with minor support.

This evaluation framework ensures that both accuracy and clinical relevance are validated. The inclusion of multi-task metrics and domain-specific interpretability analyses positions the second model as a more comprehensive diagnostic tool, forming the foundation for further research in explainable multi-modal dementia detection.

4) Problem of Biases: Biases and ethical concerns could rise from these dataset. Both the Pitt Corpus from the DementiaBank and the ADReSSo dataset present inherent biases that influence the validity and generalizability of dementia-speech models. The Pitt Corpus is heavily biased toward elderly, American English–speaking participants, recorded in clinical interview environments. Because dementia predominantly affects older adults, the dataset is not demographically representative; as a result, models trained on Pitt may inadvertently learn characteristics associated with ageing, e.g., slower speech, vocal changes, rather than cognitive impairment itself. In contrast, the ADReSSo dataset attempts to reduce demographic bias by age and gender in its dementia and control groups. However, ADReSSo still remains limited by small sample size and linguistic bias, as recordings are primarily in English and follow a structured picture-description task. This lack of linguistic, task, and environmental diversity means both datasets risk producing models that perform well under controlled research settings, but fail to generalize to broader populations, spontaneous speech, or real-world clinical use. As a matter of fact, there can be cases where people without dementia speak in a way which people with dementia usually do. This is why even though the model from the 2nd experiment was built on the ADReSSo dataset I will building it with the Pitt Corpus and see the outcomes of it.

B. Objective 2: Identify the 3 best-performing models

For the first experiment, the text-only model achieved the highest accuracy, AUC, and precision in the non-augmented dataset, confirming that linguistic modality was the most informative for this dataset. For the augmented version, however, the audio, text, and time model outperformed all other models, showing the enhancements of having all the features when using synonym replacement. Very similar placements were achieved when comparing the original version with the augmented version datasets. For the second experiment, the prediction-only model outperformed the other model, but as already stated above, the MMSE and the MMSE and Prediction model performed very badly. Which was why the use of MMSE was omitted for this research study.

As a result, the evaluation to choose the best model was performed by choosing three models, which although are different in terms of the architecture used, they use similar features and dataset types.

The second experiment is thus recommended as the best-performing, with scalability and clinical depth. Therefore, from the second research paper by the D. Altinok [25], the model using the prediction only will be our main model. On the other hand, from the first research paper used, the model using text only and the model using audio and text, will also be investigated by trying to find the best layer and buildings of these two model architectures.

I will not be using the augmented models, due to the reason that I saw them to be very prone to errors; it is easy for a person used in the training dataset to end up having their augmented version in the testing and validation dataset.

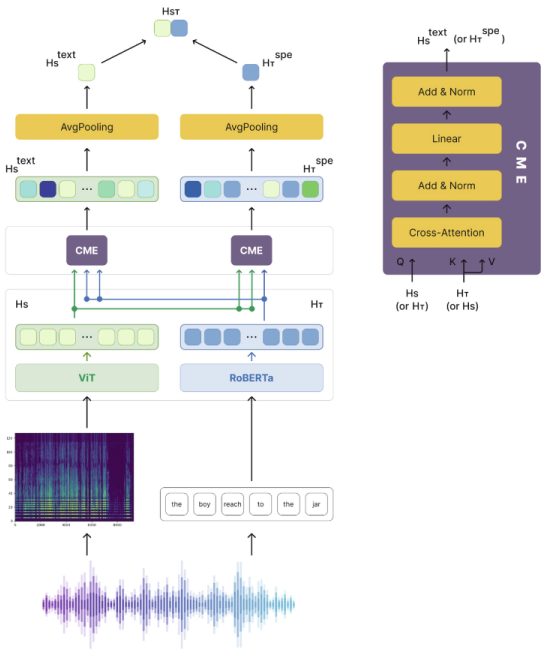

Fig. 3. Architecture design of the chosen model for the cognitive impairment screening [25].

The dataset that is used is the Pitt Corpus. I saw that the Pitt corpus is a more easily manageable dataset containing more data points and the dataset has been noise specifically removed. The Pitt corpus contains recordings of the same people over time, but not the same recording used twice, which makes it a better choice for this experiment [33].

C. Objective 3: Building a model for Dementia Classification

At this stage, for every step taken, there needs to be a reason and a strategy why I am doing so, based on previous results, findings, or information which I have learned during this semester about CNNs. Therefore, for every step taken, I will be evaluating or backing it by giving a reason.

D. Objective 4: Noting down the findings and evaluation about the final model results

As a level of evaluation, I will be using precision, recall, f1-score, accuracy, and the macro average and weight average for each of the metric used. I will also be plotting a proper confusion matrix. These metrics were chosen, since they were used by D. Altinok [25] in here study. Additionally, I will create a table summarizing all the obtained results. This will be very similar to the one used in these studies to keep the same research strategy.

VII. CONCLUSION

This report presents research and planning done to accomplish the FYP. The report has outlined the problem risen, in the creative and AI scene, whilst providing a solid understanding of the current research and information available throughout the web about dementia. This is done through multiple

research projects, practical use of models, and experiments throughout the whole summer. The last three months were used to finalize this experimentation and discuss the next steps that should be taken, which also includes the DIT meeting. As stated in GDPR article 22 [34], the final prediction when used in an application would need to be rediscussed with a psychologist, so that the decision is not solely on the model's prediction. Therefore, although this model will be built, once it is deployed in an application, it needs to be confirmed with a psychologist.

REFERENCES

[1] "Origin and history of dementia," www.etymonline.com. [Online]. Available: https://www.etymonline.com/word/dementia

[2] A. Haider, B. C. Spurling, and J. C. Sánchez-Manso, "Lewy body dementia," Nih.gov, 02 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482441

[3] N. Custodio, R. Montesinos, D. Lira, E. Herrera-Pérez, Y. Bardales, and L. Valeriano-Lorenzo, "Mixed dementia: A review of the evidence," *Dementia & Neuropsychologia*, vol. 11, p. 364–370, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5769994/

[4] S. Dattola, A. Ielo, G. Varone, A. Cacciola, A. Quartarone, and L. Bonanno, "Frontotemporal dementia: a systematic review of artificial intelligence approaches in differential diagnosis," *Frontiers in Aging Neuroscience*, vol. 17, 04 2025.

[5] K. Rascovsky, et al, "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia," *Brain*, vol. 134, pp. 2456–2477, 08 2011.

[6] H. Wei, A. V. Masurkar, and N. Razavian, "On gaps of clinical diagnosis of dementia subtypes: A study of alzheimer's disease and lewy body disease," *Frontiers in Aging Neuroscience*, vol. 15, 03 2023.

[7] GBD 2019 Dementia Forecasting Collaborators, "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*, vol. 7, 01 2022. [Online]. Available: https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(21)00249-8/fulltext

[8] C. Scerri, "Malta's strategic vision for a national dementia policy," *International Journal on Ageing in Developing Countries*, vol. 1, pp. 133–142, 2016. [Online]. Available: https://inia.org.mt/wp-content/uploads/2017/01/1.2-6-Malta-133-to-142-Final.pdf

[9] A. Scerri and C. Scerri, "Dementia in malta: new prevalence estimates and projected trends," *Original Article Malta Medical Journal*, vol. 24, 2012. [Online]. Available: https://www.um.edu.mt/library/oar/bitstream/123456789/1137/1/2012.Vol24.Issue3.A5.pdf

[10] S. Chen et al, "The global macroeconomic burden of alzheimer's disease and other dementias: estimates and projections for 152 countries or territories," *The Lancet Global Health*, vol. 12, pp. e1534–e1543, 09 2024.

[11] L. Ilias and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in Aging Neuroscience*, vol. 14, 03 2022.

[12] Z. Jahan, S. B. Khan, and M. Saraee, "Early dementia detection with speech analysis and machine learning techniques," *Discover sustainability*, vol. 5, 04 2024.

[13] L. Hung et al, "The benefits of and barriers to using a social robot paro in care settings: a scoping review," *BMC Geriatrics*, vol. 19, 08 2019. [Online]. Available: https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-019-1244-6

[14] A. Scerri and C. Scerri, "Dementia in malta : new prevalence estimates and projected trends," *Malta Medical Journal*, vol. 24, pp. 21–24, 01 2012.

[15] C. Scerri, "Empowering change: A national strategy for dementia in the maltese islands 2015-2023," Government of Malta, Tech. Rep., 01 2015. [Online]. Available: https://www.um.edu.mt/library/oar/handle/123456789/27845

[16] R. Johnson, M. M. Li, A. Noori, O. Queen, and M. Zitnik, "Graph artificial intelligence in medicine," *Annual Review of Biomedical Data Science*, vol. 7, 05 2024.

[17] S. S. Mehdoui, A. Bouzid, D. Sierra-Sosa, and A. Elmaghraby, "Dementia insights: A context-based multimodal approach," arXiv.org, 2025. [Online]. Available: http://arxiv.org/abs/2503.01226

[18] J. Hao et al, "Early detection of dementia through retinal imaging and trustworthy ai," *npj Digital Medicine*, vol. 7, 10 2024. [Online]. Available: https://www.nature.com/articles/s41746-024-01292-5

[19] R. Li, X. Wang, K. Lawler, S. Garg, Q. Bai, and J. Alty, "Applications of artificial intelligence to aid early detection of dementia: A scoping review on current capabilities and future directions," *Journal of Biomedical Informatics*, vol. 127, p. 104030, 02 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046422000466

[20] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. McWhinney, "Dementiabank," Talkbank.org, 2021. [Online]. Available: https://talkbank.org/dementia/ADReSSo-2021/index.html

[21] J. Becker and F. Boller, "Dementiabank english pitt corpus," Talkbank.org, 2025. [Online]. Available: https://talkbank.org/dementia/access/English/Pitt.html

[22] A. Bandyopadhyay, S. Ghosh, M. Bose, A. Singh, A. Othmani, and K. Santosh, "Alzheimer's disease detection using ensemble learning and artificial neural networks," *Communications in computer and information science*, vol. 1704, pp. 12–21, 01 2023.

[23] J. T. Becker, "The natural history of alzheimer's disease," *Archives of Neurology*, vol. 51, p. 585, 06 1994.

[24] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, pp. 407–422, 10 2015. [Online]. Available: https://www.cs.toronto.edu/~kfraser/Fraser15-JAD.pdf

[25] D. Altinok, "Explainable multimodal fusion for dementia detection from text and speech," in *Text, Speech, and Dialogue*, E. Nöth, A. Horák, and P. Sojka, Eds. Cham: Springer Nature Switzerland, 2024, pp. 236–251.

[26] Hugging Face, "Wav2vec2," huggingface.co, 02 2021. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/wav2vec2

[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460. [Online]. Available: https://arxiv.org/abs/2006.11477

[28] Wikipedia Contributors, "Word2vec," Wikipedia, 04 2019. [Online]. Available: https://en.wikipedia.org/wiki/Word2vec

[29] TensorFlow, "tf.keras.layers.embedding — tensorflow core v2.9.1," TensorFlow, 09 2025. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding

[30] "Vision transformer," Wikipedia, 08 2023. [Online]. Available: https://en.wikipedia.org/wiki/Vision_transformer

[31] K. Lin and P. Y. Washington, "Multimodal deep learning for dementia classification using text and audio," *Scientific Reports*, vol. 14, 06 2024.

[32] "Roberta," huggingface.co. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/roberta

[33] C. Li, W. Xu, T. Cohen, M. Michalowski, and S. Pakhomov, "Trestle: Toolkit for reproducible execution of speech, text and language experiments," *AMIA Summits on Translational Science Proceedings*, vol. 2023, p. 360, 06 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10283131/

[34] "General data protection regulation (gdpr) – legal text," https://gdpr-info.eu/, 2016, accessed: 2025-11-07.

Appendix:
Generative AI
To explore potential dataset bias, a preliminary qualitative assessment was performed with the assistance of an AI language model. The model was used to generate hypotheses about possible sources of bias in the corpora based on dataset descriptions from the literature. This exercise suggested that evaluating cross-dataset generalization may help identify overfitting to a single corpus, which motivated the methodological decision to train the model on one dataset and test it on another. This evaluation will be carried out empirically during experimentation ChatGPT link: https://chatgpt.com/g/g-p-68c1e9e95f6c8191bda398c35f67185b-thesis/shared/c/69047f0a-f2a0-832b-899c-13fc267b145f?owner_user_id=user-18bTsSpmggM4x6tyvpRSQofo.