Proceedings Paper:
Young, S., Tao, F., Mirheidari, B. et al. (2025) Can speech accurately detect depression in patients with comorbid dementia? An approach for mitigating confounding effects of depression and dementia. In: Scharenborg, O., Oertel, C. and Truong, K., (eds.) Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2025. Interspeech 2025, 17-21 Aug 2025, Rotterdam, The Netherlands. International Speech Communication Association (ISCA), pp. 499-503. ISSN: 1990-9772. EISSN: 1990-9772.

https://doi.org/10.21437/Interspeech.2025-933

Can Speech Accurately Detect Depression in Patients With Comorbid Dementia? An Approach for Mitigating Confounding Effects of Depression and Dementia

Sophie Young[1], *Fuxiang Tao*[*2], *Bahman Mirheidari*[2], *Madhurananda Pahar*[2], *Markus Reuber*[1], *Heidi Christensen*[2]

[1]School of Medicine and Population Health, University of Sheffield, United Kingdom
[2]School of Computer Science, University of Sheffield, United Kingdom

{syoung6, f.tao, b.mirheidari, m.pahar, m.reuber, heidi.christensen}@sheffield.ac.uk

Abstract

Approximately 15.9% of people living with dementia experience co-occurring major depressive disorder. Both disorders cause similar early clinical symptoms in older people but treatment options and patient outcomes differ. While it is challenging, it is therefore critical for clinicians to be able to distinguish between them. We build on existing research into objective markers of depression in speech, testing their generalizability to a more complex population. On a novel, comorbidity dataset, we demonstrate that existing depression classification methods perform worse for participants with dementia than they do for those with no cognitive decline. We also propose a method of applying Wasserstein distance-based weight vectors to emphasize depression-related information which is robust against the effect of dementia. This improves performance for users with dementia, without requiring changes to the model architectures. Our best performing model achieves an overall F1-score of 81.0%.

Index Terms: depression detection, dementia, computational paralinguistics

1. Introduction

Research into speech-based biomarkers has demonstrated that it is possible to detect a range of conditions from a person's speech including dementia [1] and depression [2]. Approximately 11.3% of people living in Europe will experience Major Depressive Disorder (MDD) in their lifetime [3]. Characterized by persistent low mood or loss of interest or pleasure, depression may be identified and diagnosed by interviews with clinicians. These interviews require a significant amount of time from a trained clinician, or rely on subjective self-reported symptoms from the patient. In view of the drawbacks of both methods, there has been considerable interest in developing more objective, automated methods to identify signs of depression or to measure its severity in speech, including the release of several publicly available speech corpora [4, 5, 6].

These automated methods do, however, often rely on the assumption that participants have depression as their sole condition, or are entirely healthy. That is, they rely on and have been evaluated on carefully controlled, homogeneous cohorts, where recordings are taken in lab settings and recruitment criteria explicitly exclude participants with comorbidities. These restrictions do not generally align with reality. In fact, depression is often found to be comorbid with a range of other health conditions including anxiety [7], cardiovascular disease [8] and dementia. On average, 15.9% of those living with dementia, another condition known to impact the articulation and content of patients' speech, also have co-occuring MDD [9]. As a result, these depression detection approaches are unlikely to perform well in a real-world population that includes the complexities introduced by a comorbid condition. This excludes many people, such as elderly people with dementia, from benefiting from these technologies.

The term dementia refers to a collection of symptoms impacting a person's cognitive abilities and daily functioning. The symptoms can be caused by several different diseases, the most common being Alzheimer's dementia. An estimated 55 million people are living with dementia, with roughly 10 million new cases annually. The challenge for a speech-based dementia detection system is that dementia and MDD often present similar early clinical symptoms, but they have different treatment options and potential patient outcomes [10]. It is, therefore, important that clinicians are able to identify whether a person presenting with symptoms such as apathy, irritability and memory complaints has dementia, MDD, or a combination of the two.

Depression and dementia are each prone to impacting the content and articulation of patients' speech. Signal processing researchers have shown significant interest in developing automated screening tools for each condition, including running challenges for both depression [4, 5] and dementia [11, 12] detection. While these disorders are investigated separately, there is considerable overlap between the methods and features used. For example, baseline models from distinct challenges for each of them both make use of the same standard acoustic feature set [5, 13]. Similarly, a decrease in speech rate and an increase in the duration and frequency of pauses have been identified as features indicative of depression [14] and dementia [15].

Given the significant overlap in methods applied, we start by testing the hypothesis that the presence of cognitive decline due to dementia will impact the performance of a depression classification model. To validate this hypothesis, we introduce a novel dataset, explained in detail in Section 3.1, which consists of speech recordings from participants with and without cognitive decline due to dementia, with a subset of each group also having depression. All recordings were collected under the same restrictions to avoid any potential bias introduced by aggregating multiple datasets. We train a Support Vector Machine (SVM) and a Long Short-Term Memory (LSTM) network, in line with previous depression classification approaches [6]. The same model architectures are trained and evaluated separately on two groups: one without cognitive decline (all participants are either healthy or have depression, that is, no comorbidity), and one where all participants have cognitive decline (meaning the depressed group has comorbidity). Both models perform worse for the participants with cognitive decline than they do for those without (Figure 1). To this end, our motivation in this work is to develop an approach to improve detection perfor-

*Corresponding author.

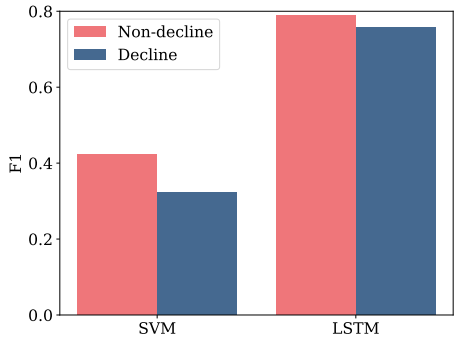

Figure 1: *F1-scores achieved on the task of depression classification for non-decline and comorbid cognitive decline groups.*

mance in comorbid depression.

We first outline a novel dataset in Section 3.1 which facilitates training a model on a mixture of speech samples with and without comorbidity, as a real-world system would encounter. We then propose a novel method of applying weights based on Wasserstein distance to the input vectors the models are trained on in order to highlight depression-relevant information and minimize the impact of comorbidities. The performance improvement is replicated in three different models. Our contributions are as follows:

- There is not, to the best of our knowledge, an existing study which aimed to develop a speech-based depression detection model that works for users with possible depression as their sole condition, as well as for those with co-occuring dementia.
- The proposed approach improves the generalizability of existing classification architectures (notably, without an increase in complexity), to improve performance for participants with cognitive decline and create a more inclusive depression classification system.

The remainder of this paper is structured as follows: Section 2 introduces related work, Section 3 introduces the proposed approach in this paper, Section 4 reports results, while Section 5 will draw conclusions.

2. Previous work

As mentioned previously, there have been several challenges aimed at developing speech-based depression screening methods [4, 5]. This has led to successful approaches such as extracting delta-mel-cepstral coefficient based features [16] (and later expanding this approach with the addition of correlation-based video features [17]) or making use of standard acoustic feature sets [18]. However, in these challenge corpora, all participants are either depressed or healthy - there is no scope for investigating the impact of comorbid conditions on depression classification.

Similarly, automated methods for detecting dementia in speech is an active research field supported by challenge datasets [11, 12]. Ensemble methods utilizing a combination of linguistic features and standard acoustic feature sets proved to be capable of dementia classification, despite possible errors introduced by automatic transcription [19]. Alternative methods leveraged pre-trained foundation models to improve classification performance [20]. Another study compared different groups of linguistic (ratios of different word classes or syntactic features), prosodic (frequency, speech rate, pitch variation) and acoustic (MFCCs, spectrum or voice quality) features [21], concluding that voice quality features are the most beneficial for dementia classification.

Given the amount of methodological overlap between depression and dementia detection, it stands to reason that the presence of one would likely interfere with the ability to detect the other automatically. Additionally, while automated detection of each of them individually has garnered significant research interest, there is comparably very little existing work exploring their interaction when co-occurring.

Some previous studies have attempted to differentiate between participants with depression and those with dementia by either collecting a single corpus that includes both groups [22] or by aggregating existing corpora [23]. However, neither study included participants who have both conditions simultaneously, and combining datasets will inherently increase the risk of introducing bias through differences in recording set up, speech elicitation tasks, etc. In the case where all participants had Alzheimer's dementia, and some also had comorbid depression, distinguishing between the two groups was a challenging task [24]. It has been shown that a support vector regressor can predict PHQ-9 scores with reasonable success in an Alzheimer's dementia population [25], but no healthy participants were included in the study. Application of transfer learning from a valency detection model can improve depression detection performance for a population with Alzheimer's dementia [26], however the effect of this on a population without cognitive decline, or a mixed population, was not investigated.

Table 1: *The demographic information in terms of gender and age. In the table, M means male while F means female.*

Group		Gender (M/F)	Age (Years)
Non-decline	Control (nD_nP)	75/92	73 ± 6
	Depressed (nD_P)	9/32	62 ± 10
Decline	Control (D_nP)	93/74	74 ± 8
	Depressed (D_P)	18/23	66 ± 10
Total	Control	168/166	73 ± 7
	Depressed	27/55	64 ± 10

3. Approach

3.1. Dataset

This study involved 416 participants: 167 reported never having or experiencing cognitive or psychiatric illness, 41 had depression only, 167 had cognitive decline only, and 41 had both conditions. The depression conditions are annotated by self-reported PHQ-9 (Patient Health Questionnaire-9), a widely used tool for assessing depression, with a score above 10 indicating depression. Cognitive decline is diagnosed using gold standard procedures by clinicians working in neurology. The speech of participants was recorded via CognoSpeak (`https://www.cognospeak.com/`) [27]. During the experiments, each participant was asked to read a short passage, "The Grandfather Passage", which uses a diverse phoneme range, helping to detect speech patterns associated with depression [28]. In this work, we merged the participants into two groups based on whether they have cognitive decline, referred to as *Decline* or *Non-decline*. Each group included individuals with and without depression. We abbreviate participants in these subgroups of

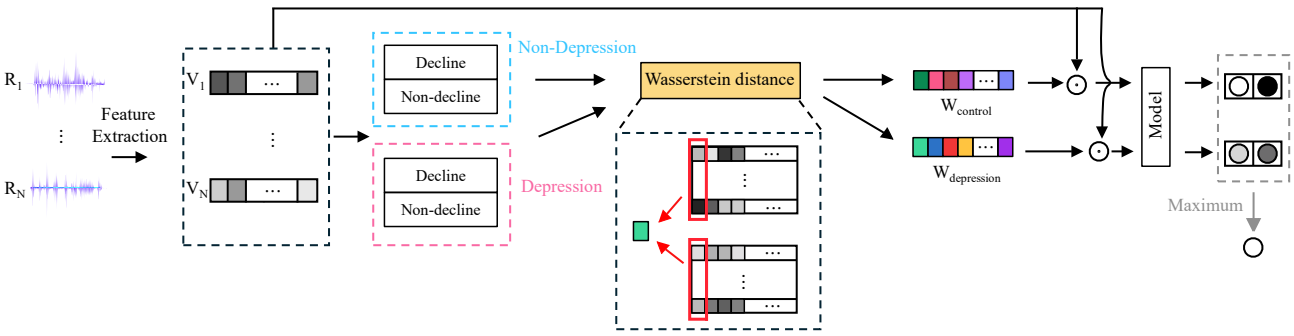

Figure 2: *The pipeline used in our proposed approach to calculate masks which emphasize depression-related information in speech.*

the *Decline* group as *D-P* (decline & depressed) and *D-nP* (decline & non-depressed), and those in the *Non-decline* as *nD-P* and *nD-nP*, respectively. This enables us to evaluate the effectiveness of our proposed approach and its ability to generalize beyond a standalone condition to a comorbid condition among participants. Table 1 provides demographic information about the participants in this dataset. There is no significant difference between *Decline* and *Non-decline* participants in terms of age distribution ($p > 0.05$ according to a two-tailed t-test) and gender balance ($p > 0.05$ according to a $\mathcal{X}^2$ test).

3.2. Experimental design

In this section, we introduce our proposed approach (shown as Figure 2) and set up a series of experiments to validate our experimental objective: namely to evaluate our proposed approach aimed at addressing the limitation of underdiagnosis of depression as a comorbid condition.

The first step of the experiment involves extracting speech features from speech signals and vectorizing them to facilitate their use in the proposed approach. To this end, we employed the well-established openSMILE toolkit [29], an open-source audio feature extraction tool widely used in affective computing. Specifically, we extracted 16 features and their delta coefficients from the Interspeech 2009 Emotion Challenge feature set [30], leading to a dimension $D = 32$ of features, which was designed for emotion recognition and mental health analysis. These features are commonly utilized to differentiate between depressed and non-depressed speech [31]. For a fair comparison, we adhered to the parameter settings of the Androids corpus [6], employing a 25 ms analysis window and 10 ms step size. After feature extraction, each recording is converted into a sequence of feature vectors $\boldsymbol{X_k} = \{x_1, x_2, x_3, ..., x_t\}$ ($k \in N$), where N is the total number of participants, and t is the total number of vectors. Since the duration of participants' recordings vary, so too does the value of t.

So far, the extracted speech features contain information relevant to differentiating between depression and non-depression, but they are influenced by cognitive decline to varying degrees, which limits the performance of models in depression detection (Figure 1); these features are used to train baseline models for this paper. The next step aims to emphasize specific depression-related information to increase robustness to the potential presence of cognitive decline, and enhance model performance in depression detection. To do this, we introduce two masks using Wasserstein distance [32], $\boldsymbol{W}_{depressed}$ and $\boldsymbol{W}_{non\_depressed}$, emphasizing features with information relating to the depressed/not-depressed condition and masking information relating to cognitive decline. The reason for using Wasserstein distance is that it can address the problem of probable inconsistent numbers of participants in the *Decline* and *Non-decline* groups within the training subset of data. The Wasserstein distance for the depressed case, $Dist_{depressed}$ is defined as follows:

$$Dist_{depressed}(a_i, b_i) = \inf_{\gamma \in \Gamma(a_i, b_i)} \mathbb{E}_{(x,y)\sim\gamma} \left[|x - y|\right], \quad (1)$$

$$W_{depressed} = softmax(\frac{1}{1 + Dist_{depressed}(a_i, b_i)}) \quad (2)$$

where a_i and b_i are the averages of the i-th feature for *D-P* and *nD-P*, respectively ($x \in a_i$, $y \in b_i$, $i \in [1, 32]$) calculated on the train set. $\Gamma(a_i, b_i)$ is the set of all possible joint distributions whose marginal distributions are a_i and b_i, and the expectation $\mathbb{E}$ is taken over the joint distribution γ. *D-P* and *nD-P* both contain depression-related information but differ in whether they include decline-related information. From this perspective, the same feature showing similarity in different groups is considered as the representation of depression-related information and will be amplified after applying the mask. This expression is written for the $\boldsymbol{Dist_{depressed}}$ and $\boldsymbol{W}_{depressed}$, but it can be used in the same way for the mask $\boldsymbol{Dist_{non\_depressed}}$ and $\boldsymbol{W}_{non\_depressed}$ using *D-nP* and *nD-nP*.

The masks $\boldsymbol{W}_{depressed}$ and $\boldsymbol{W}_{non\_depressed}$ were calculated from the training set. Then each feature vector $\boldsymbol{X}_k$ in the training set was weighted before being fed to the model, defined with the following:

$$\boldsymbol{X}_k^* = \boldsymbol{X}_k \odot (1 + W) \quad (3)$$

where W is either $\boldsymbol{W}_{depressed}$ or $\boldsymbol{W}_{non\_depressed}$, depending on the annotation of $\boldsymbol{X}_k$. At the stage of inference, each unseen feature vector in the test set is separately weighted by $\boldsymbol{W}_{depressed}$ and $\boldsymbol{W}_{non\_depressed}$. Both weighted representations are then classified, and the final prediction is determined by selecting the result where the predicted probability is furthest from the decision boundary between classes.

For a fair comparison, we utilised the same models with the same configuration in the Androids corpus [6] to evaluate the effectiveness of our proposed approach. As presented in [6], the features were averaged per recording before being passed to the SVM model, and were segmented into 32×128 frames for the LSTM model. All the models were trained with k-fold ($k = 5$) cross-validation, where all participants were split into disjoint subsets, with a different fold serving as the test set in each validation. The number of hidden states in LSTMs was set to 32. The batch size was set to 16, the learning rate was set to

Table 2: *Analysis of the effectiveness of the proposed approach in the Non-decline and Decline group. Acc, Prec, Rec and F1 represent accuracy, precision, recall and F1-score, respectively.*

Models	Group	Acc.	Prec.	Rec.	F1
SVM [6]	Non-decline	66.3	32.5	65.9	43.5
	Decline	65.4	28.8	51.2	36.8
SVM [6]+ours	Non-decline	66.3	32.9	68.3	**44.4**
	Decline	67.3	31.0	53.7	**39.3**
LSTM [6]	Non-decline	78.7	72.0	96.3	82.4
	Decline	68.9	66.7	68.7	67.7
LSTM [6]+ours	Non-decline	73.1	70.4	93.0	80.1
	Decline	67.1	66.4	77.8	**71.7**
MLA [35]	Non-decline	75.3	73.4	86.4	79.4
	Decline	68.2	67.9	80.2	73.6
MLA [35]+ours	Non-decline	81.1	80.5	90.9	**85.4**
	Decline	73.4	71.4	79.2	**75.1**

0.001 and the number of training epochs was set to 100. The training was performed with Adam optimiser [33] with categorical cross-entropy as a loss function [34].

4. Results

In this section, we present the results of our proposed method across three scenarios: i) when depression is treated as a sole condition, ii) when it appears as a comorbid condition, and iii) when both cases are considered together. In addition, we further evaluate the effectiveness of our approach by comparing it with an alternative method (MLA) from previous work [35], as this represents the current state-of-the-art method on the read task which constitutes the baseline for this paper [6].

Table 2 presents results for the first two scenarios - depression as a sole condition (*Non-decline*) and depression as a comorbid condition (*Decline*). Our proposed approach improves the detection performance for the non-decline group in SVM and MLA models but does not show improvement over the LSTM. This indicates that our approach is comparable to existing approaches when depression is considered as a sole condition. In addition, as shown in Table 2 when the best-performing models from the scenario of depression as a sole condition are applied to the scenario of depression as a comorbid condition, their performance always improves with our proposed approach, yielding F1-score improvement of 2.5%, 4%, and 1.5% for SVM, LSTM, and MLA, respectively. This demonstrates that our approach enhances the model's ability to detect depression in populations experiencing cognitive decline.

Furthermore, when combining the two scenarios above, i.e., taking both those with depression as a sole condition and those with depression as a comorbid condition into account, our proposed approach continues to improve performance across different models, shown in Table 3. In particular, our approach increases the F1-score by 1.7%, 0.2%, and 4.4% for SVM, LSTM, and MLA, respectively, reducing the average error rate by 8.9%. More importantly, such continuous improvement in the F1-score is primarily driven by stable enhancement in recall, which improves by 2.5%, 2.3%, and 2.4% for SVM, LSTM, and MLA, respectively. Higher recall reduces the likelihood of depression cases being underdiagnosed, which is particularly crucial in real-world applications. While a false positive case may only require additional examination, an underdiagnosis can result in a patient losing the optimal window for treatment, potentially leading to worse treatment outcomes and a lower quality of life.

Table 3: *Comparisons of results between our proposed approach and previous studies for the entire, combined, dataset. Acc, Prec, Rec and F1 represent accuracy, precision, recall and F1-score, respectively.*

Models	Acc.	Prec.	Rec.	F1
SVM [6]	65.9	30.8	58.5	40.3
SVM [6]+ours	66.8	32.1	61.0	**42.0**
LSTM [6]	73.8	69.8	83.0	75.8
LSTM [6]+ours	70.0	68.5	85.3	**76.0**
MLA [35]	71.9	70.7	83.4	76.6
MLA [35]+ours	77.4	76.6	85.8	**81.0**

The application of masks enhances performance on more complex datasets that include participants with comorbid dementia. Additionally, such an improvement is demonstrated across both traditional machine learning and neural network methods. This approach does not increase the complexity of the original model because it improves performance exclusively by weighting the training data without modifying the model architecture or introducing additional learnable parameters. As a result, it can be easily generalized to any established method. Moreover, the approach consistently enhances performance for participants with comorbid dementia, promoting a more inclusive classification process that is less affected by the presence of a second co-occurring disease.

5. Conclusions

In conclusion, some common baseline methods for automated depression detection have limited capacity to cope with the additional complexities of a second, comorbid, condition. However, we have shown that it is possible to increase performance for users with cognitive decline without changing the models' architectures. Our proposed method calculates and applies a mask, based on Wasserstein similarity, to the input data in order to emphasize depression-related information, which is robust against the influence of cognitive decline. Utilizing these similarity masks achieved an F1-score of 81.0% for the combined population of participants with and without cognitive decline.

One limitation of this work is that there is a significant difference in age between the depressed and non-depressed participant groups. This difference is reflected in both the non-decline and decline groups (i.e, there is no significant difference in age between *D-P* and *nD-P*, but there is between *D-P* and *D-nP*), but it is possible that this will have influenced the depression classification task. Additionally, this study groups together several different disorders under umbrella terms like "dementia" or "cognitive decline". However, it is likely that the different disorders included under the term "dementia", such as Alzheimer's disease or Lewy body dementia will impact patients' speech differently. These differences in speech pathology will also affect automatic depression detection through speech differently. Future work should consider a corpus with a more fine-grained view of dementia, allowing for investigation of differences in symptomatology.

6. Acknowledgements

This work was supported by the Sheffield BRC and NIHR funding.

7. References

[1] X. Qi, Q. Zhou, J. Dong, and W. Bao, "Noninvasive automatic detection of alzheimer's disease from spontaneous speech: a review," *Frontiers in Aging Neuroscience*, vol. 15, p. 1224723, 2023.

[2] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bio-acoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, p. 105020, 2023.

[3] L. Gutiérrez-Rojas, A. Porras-Segovia, H. Dunne, N. Andrade-González, and J. A. Cervilla, "Prevalence and correlates of major depressive disorder: a systematic review," *Brazilian Journal of Psychiatry*, vol. 42, pp. 657–672, 2020.

[4] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 2014, pp. 3–10.

[5] F. Ringeval et al., "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[6] F. Tao, A. Esposito, and A. Vinciarelli, "The androids corpus: A new publicly available benchmark for speech based depression detection," *Interspeech*, pp. 4149–4153, 2023.

[7] N. H. Kalin, "The critical relationship between anxiety and depression," pp. 365–367, 2020.

[8] A. Halaris, "Comorbidity between depression and cardiovascular disease," *International Angiology*, vol. 28, no. 2, p. 92, 2009.

[9] M. S. Asmer, J. Kirkham, H. Newton, Z. Ismail, H. Elbayoumi, R. H. Leung, and D. P. Seitz, "Meta-analysis of the prevalence of major depressive disorder among older adults with dementia," *The Journal of clinical psychiatry*, vol. 79, no. 5, p. 15460, 2018.

[10] S. Tetsuka, "Depression and dementia in older adults: a neuropsychological review," *Aging and disease*, vol. 12, no. 8, p. 1920, 2021.

[11] S. Luz et al., "An overview of the ADReSS-M signal processing grand challenge on multilingual Alzheimer's dementia recognition through spontaneous speech," *IEEE Open Journal of Signal Processing*, 2024.

[12] F. Tao et al., "Early dementia detection using multiple spontaneous speech prompts: The PROCESS challenge," *ICASSP*, 2025.

[13] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," in *ICASSP*, 2023, pp. 1–2.

[14] C. Mijnders, E. Janse, P. Naarding, and K. P. Truong, "Acoustic characteristics of depression in older adults' speech: The role of covariates," in *Interspeech*, 2023, pp. 4159–4163.

[15] Z. Liu, Z. Guo, Z. Ling, S. Wang, L. Jin, and Y. Li, "Dementia detection by analyzing spontaneous mandarin speech," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2019, pp. 289–296.

[16] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 41–48.

[17] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 2014, pp. 65–72.

[18] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 2019, pp. 81–88.

[19] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of alzheimer's disease using spontaneous speech only," in *Interspeech*, vol. 2021. NIH Public Access, 2021, p. 3830.

[20] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection," in *Interspeech*, vol. 2021. NIH Public Access, 2021, p. 3790.

[21] R. He et al., "Automated classification of cognitive decline and probable alzheimer's dementia across multiple speech and language domains," *American Journal of Speech-Language Pathology*, vol. 32, no. 5, pp. 2075–2086, 2023.

[22] B. Sumali, Y. Mitsukura, K.-c. Liang, M. Yoshimura, M. Kitazawa, A. Takamiya, T. Fujita, M. Mimura, and T. Kishimoto, "Speech quality feature analysis for classification of depression and dementia patients," *Sensors*, vol. 20, no. 12, p. 3599, 2020.

[23] M. Ehghaghi, F. Rudzicz, and J. Novikova, "Data-driven approach to differentiating between depression and dementia from noisy speech and language data," in *Proceedings of 8th Workshop on Noisy User-generated Text*, 2022, p. 24.

[24] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in alzheimer's disease through analysis of speech and language," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 1–11.

[25] D. Attas, B. Mirheidari, D. Blackburn, A. Venneri, T. Walker, K. Harkness, M. Reuber, C. Blackmore, and H. Christensen, "Predicting levels of depression and anxiety in people with neurodegenerative memory complaints presenting with confounding symptoms," in *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 3*. Springer, 2021, pp. 58–69.

[26] P. A. Pérez-Toro, D. Rodríguez-Salas, T. Arias-Vergara, S. P. Bayerl, P. Klumpp, K. Riedhammer, M. Schuster, E. Nöth, A. Maier, and J. R. Orozco-Arroyave, "Transferring quantified emotion knowledge for the detection of depression in alzheimer's disease using forestnets," in *ICASSP*, 2023, pp. 1–5.

[27] M. Pahar, F. Tao, B. Mirheidari, N. Pevy, R. Bright, S. Gadgil, L. Sproson, D. Braun, C. Illingworth, D. Blackburn *et al.*, "Cognospeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech," *arXiv preprint arXiv:2501.05755*, 2025.

[28] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in *2013 IEEE international conference on body sensor networks*, 2013, pp. 1–6.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[30] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," *Interspeech*, 2009.

[31] F. Tao, A. Esposito, A. Vinciarelli *et al.*, "Spotting the traces of depression in read speech: An approach based on computational paralinguistics and social signal processing." in *Interspeech*, 2020, pp. 1828–1832.

[32] L. N. Vaserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] R. Rubinstein and D. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*. Springer, 2004.

[35] F. Tao, X. Ge, W. Ma, A. Esposito, and A. Vinciarelli, "Multi-local attention for speech-based depression detection," in *ICASSP*, 2023, pp. 1–5.