From Pen to Prediction: Handwriting-Based Alzheimer's Detection

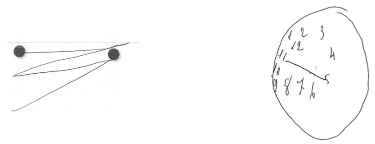
Maria Boumpi*, Kalliopi V. Dalakleidi*†, John Pavlopoulos*†
*Department of Informatics, Athens University of Economics and Business, 76 Patission Street, Athens, 104 34, Greece
†Archimedes/Athena Research Center, a Artemidos Street, Marousi, 15125, Greece

Abstract—**Early diagnosis of Alzheimer's disease (AD) is essential for timely intervention and effective care. This paper examines handwriting analysis as an accessible and non-invasive way of early detection by focusing on the comparison of raw handwriting images and tabular features. Two data sets were considered, the DARWIN handwriting dataset, which contains raw images and tabular data from pen movement, and the Alzheimer's disease dataset (ADD), which presents the patient's history and cognitive assessments in tabular format. A range of classification methods including machine learning (Random Forest, SVM, and XGBoost) was tested on tabular data from the two datasets, a deep learning Swin Transformer for image classification, and a multimodal approach that integrated both. Random Forest outperformed other models on DARWIN tabular data (83.03% ± 1.18), while XGBoost was the best on ADD (83.53% ± 3.44). The Swin Transformer also performed consistently on handwriting images (80.02% ± 0.87), capturing features associated with stroke tremors and fluency, as well as other visual aspects of the dataset. A late fusion model incorporating both modalities achieved the highest overall accuracy of 89.15% ± 1.73, showing that the image and the tabular features produce a complementary diagnostic value. These results indicate that handwriting includes fine neuromotor features related to early AD that can surpass clinical conventional data. We also present ablation studies on task order with respect to image training and end-to-end multimodal learning. These findings provide further evidence of the benefits of modular fusion in situations where data is restricted. Handwriting samples have the potential to become a useable and scalable resource in AD screening because of their low cost, ease of collection, and acquisition logistics, which even accommodate home-based settings.**

I. Introduction

Progressive neurodegenerative diseases (ND), such as Alzheimer's disease (AD), affect millions of individuals worldwide. As the most common form of dementia, AD cases involve devastating memory loss, decline of cognitive and motor functions, and loss of daily function [1]. Thus, early diagnosis is critical to slow progression and improve the efficiency of subsequent treatment [2].

Handwriting is a complex activity that involves cognitive, motor and perceptual processes, making it a promising marker of cognitive impairment [3], [4]. In AD, patients often display disturbances in spatial organization, movement control, and stroke consistency [4]. Digital tablets allow for a systematic study of these effects by collecting raw handwriting images and fine-grained kinematic features such as pen pressure, velocity, and spatial coordinates (see Figure 1).



(a) Task 2 – Join two points with a horizontal line

(b) Task 24 – Draw a clock showing 11:05

Fig. 1: Examples of AD handwriting from the DARWIN dataset [3]

Clinical datasets cover an extensive medical history, but subtle handwriting motor signs capture changes before the onset of clinical symptoms. There is a critical gap in our understanding of which data type will yield the most value in diagnosis: clinical, tabular features, raw handwriting images, or their combination. The tabular features of handwriting are well described, while visual representations remain underexplored despite their potential to reveal fine-grained patterns such as tremor or stroke fluidity.

In our study, we combine modalities and demonstrate the added diagnostic value handwriting images offer, especially, paired with tabular attributes for early-stage diagnosis. It investigates the following research question: *Can visual handwriting signals and motion-derived features result in more effective early Alzheimer's diagnosis compared to traditional medical history information?* To address this question, we focus on three main contributions: (i) analyze the effectiveness of clinical, tabular, and image-based handwriting features; (ii) present a fusion model where handwriting images and handcrafted tabular features are integrated; (iii) show that late fusion significantly improves performance and robustness by leveraging the complementary strengths of both modalities.

II. Related Work

The DARWIN dataset has supported various machine learning approaches for Alzheimer's detection, using temporal and spatial handwriting features. Baseline studies applied support vector machines (SVM), random forest (RF) and k-nearest-neighbors (KNN) [3], while subsequent studies added ensemble models with SHAP and highlighted key features such as pen pressure and in-air time [5]. Moving beyond tabular attributes, new hybrid models incorporate time-series data and reconstructed handwriting images [6], while CNNs are also used to analyze grayscale pen data [7]. Digital drawing tests such as DCTclock have also shown superior sensitivity to mild cognitive impairment compared to Mini Mental State Examination (MMSE) [8]. Unlike these unimodal studies, our study combines DARWIN tabular features and handwriting images and demonstrates that integration of both modalities enhances stability and boosts classification efficacy.

III. METHODS

A. Datasets

The DARWIN dataset [1] was made for early AD detection via handwriting analysis and consists of data from 174 subjects (89 AD and 85 healthy). Each participant has 452 features: 1 ID, 1 class label, and 450 kinematic features from 25 writing tasks encompassing graphic, copying, memory, and dictation activities. 18 temporal and spatial features were extracted from each task, including time, speed, jerk, pressure, motion metrics, and jerk, covering various parameters of handwriting motion. Handwriting images are available for six tasks (2, 3, 4, 5, 21, and 24), with almost full coverage for AD participants (88 out of 89) and somewhat less for healthy controls (78 out of 85) likely due to consent limitations.

The Alzheimer's Disease Dataset (ADD) [9] is available on Kaggle, and contains clinical and demographic data for 2,149 individuals (760 with AD, 1,389 without AD). Each record contains 34 attributes, including demographic, lifestyle, and clinical data. A total of 32 attributes (12 numerical, 20 categorical/binary) were retained after preprocessing. ADD also contains cognitive (e.g. MMSE) and functional (e.g. Activity of Daily Living (ADL)) assessments and symptoms reports such as memory complaints, confusion, and disorientation.

B. Data Preprocessing

To achieve consistency across modalities, the DARWIN tabular dataset was filtered to include only image data tasks (2, 3, 4, 5, 21, 24). We kept only those participants who had images, resulting in 88 AD patients and 78 healthy controls with complete multimodal data. For the ADD dataset, we subsampled 166 participants (88 AD, 78 healthy), to achieve a balanced design that matched the size and class distribution of DARWIN. Although the two data sets contain different individuals, this alignment allowed parallel analyses in similar experimental settings. At the participant level, data splits for tabular and image experiments were synchronized. For each random seed, matching IDs were assigned to the training and test sets in order to achieve a fair comparison and reliable fusion.

C. Tabular Data Classification

For ADD and DARWIN (tabular), multiple binary classifiers were trained to distinguish AD patients from healthy controls. For this purpose, six standard machine learning models were used: (1) Support Vector Machine (SVM), (2) Logistic Regression (LR), (3) Random Forest (RF), (4) Gaussian Naive Bayes (GNB), (5) k-Nearest Neighbors (KNN), and (6) XGBoost (XGB). These models were chosen considering their performance on binary classification, probability outputs, and feature importance.

Each model was evaluated using five Monte Carlo cross-validation runs (80% training and 20% testing), the results were averaged and Standard Error of the Mean (SEM) was reported to show variability. Hyperparameter tuning was performed by grid search, Optuna, and default settings were used to ensure fair and consistent comparisons across models while exhaustively searching the parameter space. Stratified sampling was applied throughout to maintain class balance.

TABLE I: Sample features from DARWIN tabular data for task 2

Feature	Value	Description
AIR_TIME2	6,085	Pen-up time (ms)
MAX_X_EXTENSION2	4,945	Max horizontal stroke (px)
PRESSURE_MEAN2	1,851.08	Mean pen pressure
TOTAL_TIME2	24,870	Task duration (ms)

TABLE II: Sample features from ADD

Feature	Value	Description
FAMILYHISTORYALZHEIMERS	0	Family history of AD
DEPRESSION	1	Diagnosed depression
FUNCTIONALASSESSMENT	6.52	Daily function score
MEMORYCOMPLAINTS	0	Memory complaint (yes/no)

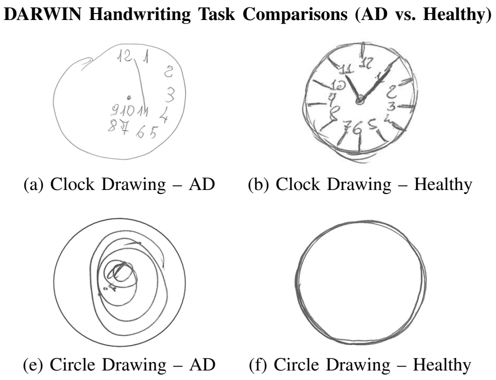
D. Image data classification

For handwriting-based AD detection, we employed a fine-tuned Swin Transformer [10], chosen for its ability to capture both local stroke details and global handwriting structure. For the handwriting images corresponding to tasks 2, 3, 4, 5, 21, and 24, we resized the images to 224×224, and normalized them with ImageNet statistics. Consistent with the tabular experiments, the data was split into 80% training/validation (divided into 72% training, 8% validation) and the remaining 20% for testing.

To improve generalization, training involved shuffling images each epoch, while the validation and test sets were kept constant to allow consistent and reproducible evaluation. Training used AdamW (learning rate = 5e-5), cosine annealing, cross-entropy loss with label smoothing ($\varepsilon = 0.1$) and early stopping (patience = 10) to avoid overfitting. We froze the lower layers and fine-tuned the deeper layers. The best model per seed was selected by validation loss and evaluated in the test set. Generalization was assessed using Monte Carlo cross-validation (seeds 42–46) and the results were reported as mean accuracy with SEM.

E. Multimodal Fusion Model

To evaluate the combined predictive power of both image-based and tabular features, we used a late fusion approach by aggregating the output probabilities of separately trained models. For the image modality, we used a fine-tuned Swin Transformer (see Section III-D to assess handwriting images



(a) Clock Drawing – AD (b) Clock Drawing – Healthy

(e) Circle Drawing – AD (f) Circle Drawing – Healthy

Fig. 2: Tasks 2 (clock) and 3 (circle) from the DARWIN dataset, comparing AD (left) and healthy (right) samples.

and generate class probabilities for each participant. For the tabular modality, we used a RF classifier trained on the corresponding hand-crafted handwriting features extracted from the same subset of DARWIN tasks. RF was selected because it achieved the highest classification accuracy among all evaluated models in tabular classification experiments (see Section IV-A).

For inference, we ensured that each model used the same participant ID and test split for each random seed. To achieve this, image filenames were processed to retrieve participant IDs, which were used to find the corresponding rows of the tabular test data. Softmax normalized class probabilities were generated by each model and the final prediction was made by averaging the two class probability distributions, late mean fusion. This method provides each modality an equitable contribution to the final prediction.

Model evaluation employed Monte Carlo cross-validation using five random seeds (42-46). The accuracy of each seed was computed and the results were summarized using mean accuracy and SEM, providing a measure of stability and generalization.

IV. RESULTS

A. Tabular data

The performance of classical machine learning classifiers was evaluated on DARWIN handwriting features and the ADD clinical dataset using five Monte Carlo cross-validation runs. The results confirm that the clinical data for ADD yielded slightly higher performance, likely due to its wider diversity of features. Nevertheless, DARWIN tabular features also demonstrated competitive accuracy, especially with optimized classifiers. In DARWIN, RF achieved the highest accuracy, while in ADD, XGB outperformed all other models. The complete results are reported in Table III.

TABLE III: Mean accuracy and Standard Error of the Mean (SEM) of classifiers on DARWIN and ADD, in bold the best

Classifier	DARWIN	ADD
Random Forest (RF)	**83.03% ± 1.18**	80.59% ± 2.56
Support Vector Machine (SVM)	79.50% ± 1.62	75.88% ± 1.10
Gaussian Naive Bayes (GNB)	78.21% ± 2.12	77.06% ± 2.85
Logistic Regression (LR)	81.29% ± 1.99	76.47% ± 3.22
k-Nearest Neighbors (KNN)	75.22% ± 2.90	74.12% ± 2.53
XGBoost (XGB)	82.77% ± 4.05	**83.53% ± 3.44**

B. Handwriting images

The Swin Transformer was evaluated on raw handwriting images from selected DARWIN tasks. It achieved an average accuracy of 80.02% (±0.87), with results across seeds ranging from 77.27% to 82.29% (Table IV). Although image-based performance did not surpass the best tabular models (e.g., RF in DARWIN at 83.03% and XGB in ADD at 83.53%), it remained strong and consistent in multiple splits. These results highlight the ability of vision transformers to extract meaningful patterns from handwriting images and demonstrate their utility as a standalone diagnostic tool when tabular features are unavailable.

TABLE IV: Classification accuracy of Swin on DARWIN handwriting images per seed (from 42 to 46) and overall.

Seed	Test Accuracy (%)
42	82.29
43	80.65
44	78.92
45	80.95
46	77.27
Mean ± SEM	**80.02 ± 0.87**

C. Fusion Results

The multimodal fusion model combining Swin Transformer image predictions with RF tabular predictions, achieved the highest overall performance. By averaging the output probabilities, the fusion strategy leveraged complementary strengths of visual and kinematic features.

As shown in Table V, it consistently outperformed unimodal models across five seeds, with a mean accuracy of 89.15% (±1.73) and individual runs ranging from 84.85% to 93.55%. This demonstrates the effectiveness of integrating visual and tabular handwriting data, with fusion surpassing both single-modality baselines and the best clinical dataset models. It underscores handwriting as a powerful and practical modality, capable of supporting scalable diagnostic systems and, in some cases, outperforming traditional clinical data.

TABLE V: Fusion model accuracy across five seeds (42–46), combining Swin Transformer image predictions with RF tabular predictions.

Seed	Test Accuracy (%)
42	90.62
43	**93.55**
44	85.29
45	91.43
46	84.85
Mean ± SEM	**89.15 ± 1.73**

V. Error Analysis

To better understand model behavior and robustness, we performed a detailed error analysis comparing the best-performing seed (43) and the worst-performing seed (46).

A. Confusion Matrices

The confusion matrices for seeds 43 and 46 are in Fig 3. For seed 43, the model provided balanced performance, correctly classified all 16 patients and 13 healthy individuals, with only one false positive and one false negative. Seed 46, on the contrary, did not produce false negatives, but misclassified five healthy individuals as patients. In medical screening, false negatives are more problematic as they indicate missed diagnoses that could delay treatment. Thus, seed 46 is advantageous in minimizing this critical error, while seed 43 offers fewer total misclassifications. Although it is preferable for a model to avoid false positives, the impact of false positives is more tolerable since they typically lead to further clinical evaluation, where specialists can rule out pathology and confirm cognitive status.

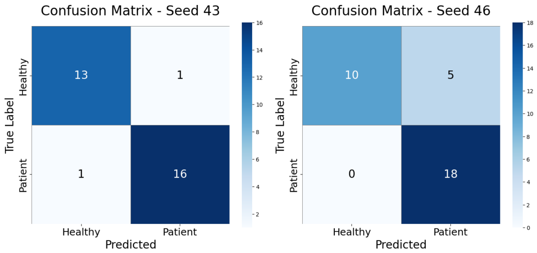


Fig. 3: Confusion matrices for seed 43 (left) and 46 (right).

B. Confidence Distribution Analysis

For both the best performing (seed 43) and the worst performing (seed 46) runs, the distribution of prediction confidence was analyzed and clear distinctions were noted. In both cases, higher confidence was associated with correct predictions and lower confidence with incorrect predictions. Seed 43 demonstrated strong calibration, correct predictions were concentrated within the 0.65–0.95 range, and errors appeared only with low confidence (below 0.6). Seed 46, however, had a greater calibration with errors at even moderate and high confidence (above 0.70) suggesting possible poor calibration and overconfidence. It was noted that both models had an understanding of confidence and its correctness, but seed 43 was more reliable and better calibrated in this respect. These results highlight the value of using confidence thresholds (e.g., 0.7) in clinical settings to flag uncertain predictions for expert review, thus reducing the risk of potentially misleading outputs.

C. The Role of Contribution Balance in Prediction Accuracy

Figure 4 shows the contributions of the Swin Transformer, the RF model, and their agreement in correct predictions across five seeds. Overall, Swin contributed more frequently than RF, especially in seeds 43–45, while seed 46 was the only exception, the RF's contribution was equal to Swin's in cases of disagreement, coinciding with the lowest fusion accuracy (84.85%). In contrast, the highest accuracy (93.55%, seed 43) aligned with a strong Swin contribution, indicating its predictions are more reliable in cases of disagreement. These findings emphasize that intermodel agreement drives most correct classifications, but when models diverge, Swin plays the more decisive role. This underscores the importance of a well-calibrated visual model within the late fusion framework.

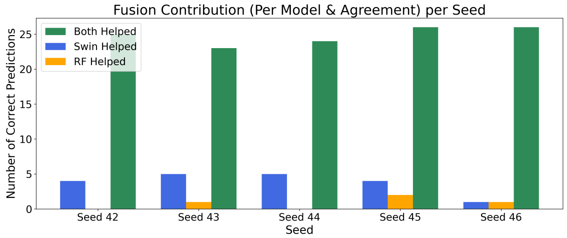


Fig. 4: Proportion of predictions where Swin or RF had higher confidence, per seed (42-46).

D. Examples Demonstrating the Role of Visual Context

To illustrate the added value of the image-based model, we investigated cases where Swin accurately predicted AD while RF failed. Figure 5 presents handwriting samples from participant `id_45`, whom RF misclassified as healthy. While Swin leveraged information across multiple tasks, we present examples from tasks 2 and 5 as they generated the highest confidence predictions. In this participant's tabular features, we noticed atypically low `pressure_var` and shorter task completion times, which statistically resembled healthy dynamics and likely contributed to RF's misclassification. In contrast, the handwriting images enabled Swin to correctly classify the case by detecting subtle spatial distortions and stroke instability. Examples like this illustrate the risk of relying solely on abstracted statistical features and highlight the value of raw visual input as evidence of detecting cognitive impairment.

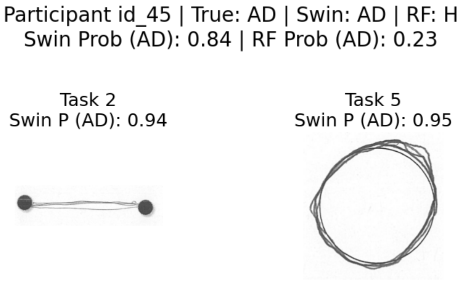

Participant id_45 | True: AD | Swin: AD | RF: H
Swin Prob (AD): 0.84 | RF Prob (AD): 0.23
Task 2
Swin P (AD): 0.94
Task 5
Swin P (AD): 0.95

Fig. 5: Tasks 2 and 5 for Participant id_45 (AD). High-confidence Swin predictions (≥ 0.94) correctly classified the case, unlike RF.

VI. Ablation Study

We conducted two ablation studies to examine fundamental design choices. First, we tested cumulative task-wise learning, where the Swin Transformer was trained sequentially across tasks instead of shuffled samples, simulating memory accumulation. The best order (21 → 24 → 5 → 4 → 3 → 2) reached an accuracy of 79.80%, slightly below the baseline, indicating that task order may influence learning dynamics. Next, we used an end-to-end multimodal model in which Swin image embeddings were concatenated with tabular feature representations from a connected neural network, and passed through a joint classification head. It achieved 79.92% accuracy, underperforming late fusion and suggesting that separate optimization per modality is more effective with limited data.

VII. Computational Resources and Latency

The traditional ML models and the fusion model were run in Python 3.9.13 with scikit-learn 1.2.2 on a workstation (Intel Core i5-10210U, 16 GB RAM). Training the Random Forest in the fusion model took 4.3 s, while the full five-run Monte Carlo cross-validation required 5 min 34 s. The Swin Transformer was trained in PyTorch 2.6.0+cpu on Lightning AI with an NVIDIA A100 GPU. Hyperparameter tuning was performed during cross-validation but is excluded from the reported time, since it is an offline process.

VIII. Discussion

Our dataset contained handwriting pictures from six simple drawing activities, including lines, spirals, and basic geometric shapes. Although useful, these tasks lacked the linguistic and recall factors known to be strong predictors of cognitive decline [11], expanding to more cognitively complex activities such as copying, recalling, or dictating sentences or words would offer more informative input. Even with this limited input, our models achieved competitive accuracy, and the fusion technique shows that well-aligned multimodal datasets of a constrained size can be highly effective.

The results likely differ in part from related work cause of the smaller scale and limited task variety in our dataset. Prior studies reported accuracies of 85–94% using larger datasets and advanced methods [3], [5], [6], [9]. The smaller image-aligned subsets reached 83.03% in DARWIN and 83.53% in ADD, strong results but still below the 89.15% of the fusion approach.

In addition to the value of the performance itself, handwriting data can be rapidly and easily collected using digital technology or scanned paper without requiring any specialized tools. Handwriting images capture diagnostic cues, such as tremors, stroke irregularities, hesitation, gaps, and reduced fluency, that may be overlooked in engineered features. Also, handwriting tasks can be administered remotely, allowing self-screening outside clinics. Consequently, handwriting based analysis can be considered accessible, scalable, and cost-efficient for the initial stages of AD detection, especially in low resource environments.

IX. Conclusion

This study explored the use of handwriting images and hand-crafted tabular features in the early detection of Alzheimer's using the DARWIN dataset. We looked at traditional machine learning, a Swin Transformer, and multimodal fusion approaches. Each modality showed efficacy, although the late fusion approach provided the best performance (89.15% ± 1.73) illustrating the benefit of integrating visual handwriting features with motion-derived information. Handwriting images often corrected misclassifications from tabular data, and ablation studies confirmed their robustness and the superiority of late fusion over joint fusion. Overall, handwriting emerges as a valuable early screening marker for Alzheimer's and can potentially outperform traditional clinical data. Its low cost, ease of use, and suitability for remote application offer a unique opportunity for scalable early detection.

X. Acknowledgement

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

[1] A As. 2023 alzheimer's disease facts and figures. *Alzheimers Dement*, 19(4):1598–1695, 2023.

[2] Bruno Dubois, Harald Hampel, Howard H Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, et al. Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323, 2016.

[3] Nicole D Cilia, Giuseppe De Gregorio, Claudio De Stefano, Francesco Fontanella, Angelo Marcelli, and Antonio Parziale. Diagnosing alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111:104822, 2022.

[4] Perla Werner, Sara Rosenblum, Gady Bar-On, Jeremia Heinik, and Amos Korczyn. Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 61(4):P228–P236, 2006.

[5] Ngoc Truc Ngan Ho, Paulina Gonzalez, and Gideon K Gogovi. Writing the signs: An explainable machine learning approach for alzheimer's disease classification from handwriting. *Healthcare Technology Letters*, 12(1):e70006, 2025.

[6] Changqing Gong, Huafeng Qin, and Mounîm A El-Yacoubi. Hybrid transformer for early alzheimer's detection: Integration of handwriting-based 2d images and 1d signal features. *arXiv preprint arXiv:2410.10547*, 2024.

[7] Pakize Erdogmus and Abdullah Talha Kabakus. The promise of convolutional neural networks for the early diagnosis of the alzheimer's disease. *Engineering Applications of Artificial Intelligence*, 123:106254, 2023.

[8] Jin-Hyuck Park. Clock drawing test with convolutional neural networks to discriminate mild cognitive impairment. *The European Journal of Psychiatry*, 38(3):100256, 2024.

[9] Rabie El Kharoua. Alzheimer's disease dataset. https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset, 2024.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[11] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.