Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2132–2142
https://doi.org/10.54941/ahfe1008439

A Task Selection Method for Alzheimer's Disease Patient Prediction to Reduce Patients' Burden

Soma Furukawa[1] and Masaomi Kimura[1,2]

[1]Shibaura Institute of Technology, Tokyo, Japan
[2]Universiti Malaysia Kelantan, Kelantan, Malaysia

ABSTRACT

Alzheimer's disease (AD) is one of the leading causes of dementia, and its diagnosis can impose a burden on patients due to examinations requiring specialized equipment and sustained effort. Handwriting-based assessment has attracted attention as a simplified diagnostic approach; however, such assessments may still require older adults suspected of having AD to perform multiple tasks. This study proposes a task selection method to reduce the number of handwriting tasks while maintaining classification performance. For each task-feature combination, measurements from 165 participants were compressed into a one-dimensional latent representation using an Autoencoder (AE). The resulting representations were organized into a task-by-feature matrix, and Principal Component Analysis (PCA) was applied to characterize the tasks. Tasks considered to contribute little to discrimination were identified based on their principal component scores and removed. Classification performance was evaluated using Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The results showed that accuracy decreased for SVM and MLP as the number of tasks was reduced, whereas RF maintained relatively stable accuracy even with fewer tasks. These results suggest that the proposed method can identify tasks that contribute little to discrimination while reducing the number of tasks required for AD prediction. However, the reduction in examination burden was inferred from the reduction in the number of tasks and was not directly evaluated in terms of task difficulty. Future work should therefore develop and validate objective indices of task difficulty to further evaluate the effectiveness of the proposed task reduction method.

Keywords: Alzheimer's disease, Handwriting tasks, Task selection, Deep learning

INTRODUCTION

According to the World Health Organization (WHO), approximately 47 million people worldwide are affected by dementia, and this number is projected to reach around 130 million by 2050 (World Health Organization, 2017). Dementia is one of the major causes of dependency among older adults, and Alzheimer's disease (AD) is its most common form, accounting for approximately 60–70% of all cases. AD initially manifests as mild symptoms such as forgetfulness and progressively impairs cognitive functions, including temporal and spatial orientation (Kindai University Hospital, n.d.). Although no definitive cure for AD currently exists, pharmacological treatments can

Received September 11, 2026; Revised October 13, 2026; Accepted October 20, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

alleviate its symptoms (Japanese Society of Neurology, 2017). Therefore, the early detection of AD is an important challenge for slowing disease progression.

One approach to the early detection of AD is to analyze handwriting and drawing characteristics using machine learning to distinguish individuals with AD from healthy individuals. The DARWIN dataset developed by Cilia et al. contains motor features, such as pen pressure and writing speed, obtained from multiple handwriting and drawing tasks, and has been used to investigate the effectiveness of handwriting-based AD classification (Cilia et al., 2018; Cilia et al., 2022). However, such assessments require participants to perform multiple tasks. Therefore, reducing the number of tasks while retaining the information necessary for classification is important for reducing the burden associated with the assessment.

Previous studies have investigated reducing the number of handwriting tasks or features while maintaining classification performance (Gattulli et al., 2023; Moreira et al., 2024). These approaches mainly selected tasks or features according to their usefulness for classification, with classification results or performance serving as the basis for determining which information should be retained.

On the other hand, differences in handwriting motor characteristics between individuals with AD and healthy individuals have also been reported (Slavin et al., 1999). In addition, previous studies have investigated the relationships between handwriting characteristics and factors such as task difficulty and mental workload (Luria and Rosenblum, 2012; Babushkin et al., 2023). These findings suggest that handwriting motor features may reflect not only the condition of the participant but also the characteristics of the task being performed.

Based on these findings, this study focuses on differences in handwriting motor features between individuals with AD and healthy individuals rather than using classification performance itself as the criterion for task selection. Tasks showing only small differences in motor characteristics between the two groups are assumed to provide relatively limited information for distinguishing them and are therefore considered candidates for removal. We then examine whether classification performance can be maintained after removing such tasks. The aim of this study is to propose a task selection method based on differences in motor characteristics and thereby reduce the number of tasks required for AD classification.

Related Work

Machine learning has been widely investigated for the detection and classification of Alzheimer's disease (AD) (Kavitha et al., 2022; Alatrany et al., 2024).

Handwriting-based approaches have been investigated as a means of classifying individuals with Alzheimer's disease (AD) and healthy individuals using machine learning. Cilia et al. constructed the DARWIN dataset by collecting motor features from individuals with AD and healthy individuals performing multiple handwriting and drawing tasks (Cilia et al., 2018;

Cilia et al., 2022). They further applied machine learning methods to these features and demonstrated the potential of handwriting and drawing information for distinguishing individuals with AD from healthy individuals.

Research has also investigated reducing the number of tasks or features used in handwriting-based AD classification while maintaining classification performance. Gattulli et al. analyzed patterns in AD classification results and investigated a method for selecting handwriting tasks that were effective for classification (Gattulli et al., 2023). Moreira et al. applied dimensionality reduction and explainability techniques to the DARWIN dataset and examined classification performance after reducing the number of features (Moreira et al., 2024). These studies reduced the amount of information used for classification by identifying tasks or features that were useful based mainly on their relationships with classification results or performance.

In parallel, the relationship between handwriting motor characteristics themselves and AD has also been investigated. Slavin et al. reported that individuals with AD exhibited lower consistency and greater variability in handwriting movements than healthy individuals (Slavin et al., 1999). Handwriting motor features have also been shown to vary according to the characteristics of the task being performed. Luria and Rosenblum investigated handwriting under different levels of mental workload and reported differences in temporal, spatial, and angular velocity measures (Luria and Rosenblum, 2012). Babushkin et al. further analyzed the relationship between handwriting task difficulty and kinematic features (Babushkin et al., 2023). These studies suggest that handwriting motor features may reflect not only the condition of the participant but also the difficulty and characteristics of the task.

However, conventional task-reduction approaches have primarily evaluated task usefulness based on classification results or classification performance. In contrast, methods that use the differences in motor characteristics between individuals with AD and healthy individuals for each task as the criterion for task selection have not been sufficiently investigated. Therefore, this study quantifies the distributions of motor features for the two groups in each task and selects tasks based on the differences between these distributions. This approach aims to provide a criterion for task selection based on motor characteristics without requiring classification performance itself as the criterion for selecting tasks.

Proposed Method

This study proposes a method for representing the distribution of measurements across subjects for each task-feature combination as a single scalar value. The proposed method uses the DARWIN dataset used in previous studies. After excluding nine healthy subjects with anomalous measurements in Task 19, a total of 165 subjects, consisting of 89 patients with Alzheimer's disease (AD) and 76 healthy controls (HC), were included in this study.

The DARWIN dataset contains 18 features measured across 25 handwriting tasks; therefore, measurements from 165 subjects are available for each of the 25×18 task-feature combinations.

First, the measurements of the 165 subjects were standardized separately for each task-feature combination to eliminate differences in scale among features. The standardized measurements were then input into an Autoencoder (AE) as a single 165-dimensional vector. The AE was trained to reconstruct this vector through a one-dimensional bottleneck, allowing the latent value to encode characteristics of the measurements across subjects. This one-dimensional latent value was used as a compact representation of the measurement distribution. This process was applied to all 25×18 task-feature combinations, resulting in 450 one-dimensional values that were arranged into a 25×18 matrix. Thus, the 18 features associated with each task were converted into numerical representations of the distribution of measurements across subjects.

Subsequently, Principal Component Analysis (PCA) is applied to the 25×18 matrix obtained using the proposed method to analyze the characteristics of each task, represented by its 18 features, in a lower-dimensional space. The meaning of the principal components is interpreted by examining the relationships between the principal component scores and the one-dimensional output values. Based on these relationships, the principal components that reflect differences in motor characteristics between individuals with Alzheimer's disease (AD) and healthy individuals are identified. The tasks are then ranked according to their scores on the relevant principal components, and tasks showing small differences in motor characteristics between the two groups are designated as candidates for removal. Finally, to determine whether the selected tasks can be removed without substantially degrading classification performance, tasks are incrementally removed according to this ranking, and the classification accuracy of supervised learning models is evaluated. Figure 1 shows the structure of the proposed AE.

In this study, task selection is performed using the measurements of all 165 subjects through unsupervised learning with an autoencoder (AE) and principal component analysis (PCA). This process does not use AD/HC class labels for learning but instead aims to characterize each task based on the distribution of measurement values. In the subsequent classification experiment, the measurements from the selected tasks are used, whereas the outputs obtained from the AE and PCA are not directly used to train the classification models.

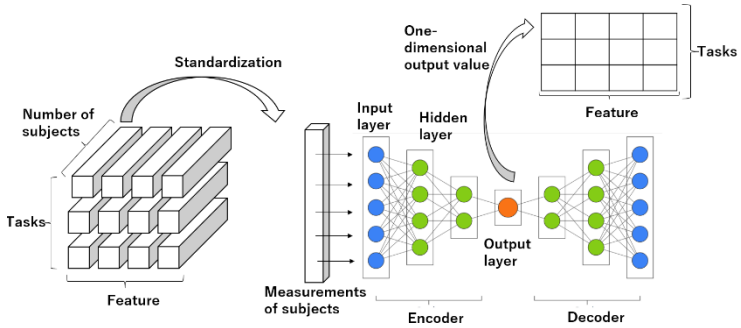


Figure 1: Model architecture for latent representation extraction using an AE.

Experiment 1: Dimensionality Reduction of Data From All Subjects Using an AE

In this experiment, we examine whether the one-dimensional latent value obtained by the AE reflects differences in the measurement distributions across the 165 subjects. The AE consists of an encoder with an input layer of 165 nodes, two hidden layers with 64 and 32 nodes, respectively, and an output layer with one node. The decoder has a symmetrical structure. The AE was trained for 1,000 epochs using the standardized measurements of the 165 subjects for each task-feature combination as input. The loss decreased during training, confirming that the training progressed appropriately.

Figure 2 shows a heatmap of 450 standardized 165-dimensional measurement vectors, sorted by the one-dimensional output values produced by the encoder. As shown in Figure 2, when the one-dimensional output value was low, the measurements of individuals with AD tended to be higher, whereas when the output value was high, the measurements of healthy individuals tended to be higher. This result indicates that the one-dimensional output value captures differences in the measurement distributions between individuals with AD and healthy individuals.

To further examine the relationship between the one-dimensional output values and the measurement distributions, integer values ranging from −68 to 50, corresponding to the observed output range, were input into the decoder at intervals of 1 to reconstruct the corresponding measurement distributions. Figure 3 shows that the reconstructed measurement distributions changed continuously with the one-dimensional output value. In addition, a boundary between the measurement distributions of individuals with AD and healthy individuals was observed near an output value of 0. At this boundary, neither group showed a clear bias.

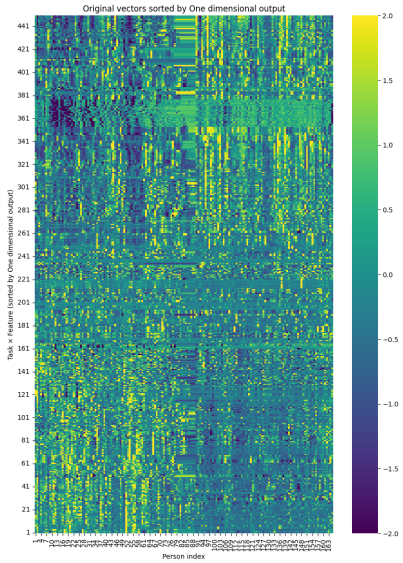


Figure 2: Heatmap of standardized measurement distributions sorted by one-dimensional output value.

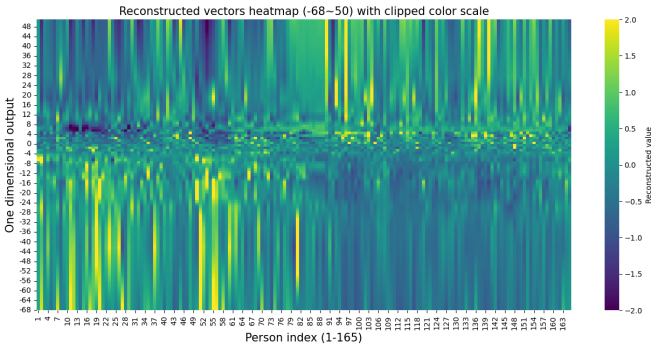


Figure 3: Heatmap of measurement distributions reconstructed by the decoder.

Experiment 2: Interpretation of Principal Components via PCA and Determination of Key Task Groups

Using the one-dimensional outputs obtained in Experiment 1, we constructed a 25×18 matrix, with tasks as rows and feature values as columns, and performed Principal Component Analysis (PCA). Because the cumulative contribution ratio of PC1 and PC2 exceeded 80%, these two principal components were selected for interpretation.

An examination of the factor loadings revealed that specific features contributed substantially to both PC1 and PC2. Therefore, we focused on the 12 features that contributed substantially to these principal components and analyzed the relationship between the distributions of the one-dimensional outputs and the principal component scores when the tasks were arranged according to their scores. The results are shown in Figure 4.

For PC1, the range of one-dimensional output values tended to shift from negative to positive values as the principal component score increased. Based on the correspondence between this tendency and the results of Experiment 1, PC1 was interpreted as representing whether the measurement values of the task features tend to be higher for individuals with AD or for healthy individuals.

In contrast, for PC2, the distribution of the one-dimensional output tended to become concentrated around zero as the principal component score increased. This suggests that a larger proportion of the features exhibited small differences in measurement values between individuals with AD and healthy individuals. Therefore, PC2 was interpreted as representing the magnitude of the differences in measurement values between individuals with AD and healthy individuals for the task features.

In this study, tasks that exhibit similar motor characteristics in individuals with AD and healthy individuals are considered candidates for removal because they are assumed to provide limited information for distinguishing between the two groups. Accordingly, we assumed that tasks with higher PC2 scores would exhibit smaller differences between the two groups and ranked the tasks according to their PC2 scores.

Furthermore, task groups were defined using PC2 scores as thresholds to compare classification performance across different levels of

task reduction. Because the distribution of PC2 scores was concentrated around and above zero, four task groups were defined using the following thresholds: ≤ -40, ≤ -20, ≤ 0, and all tasks. The numbers of tasks included in these groups were 4, 6, 10, and 25, respectively. The resulting task groups are shown in Figure 5.

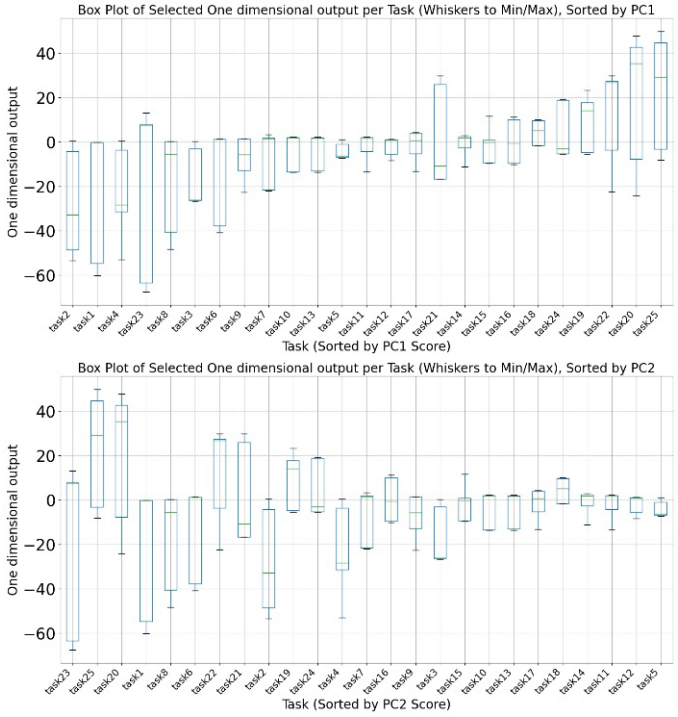


Figure 4: Distributions of 12 one-dimensional output values for tasks sorted by PC1 and PC2 scores.

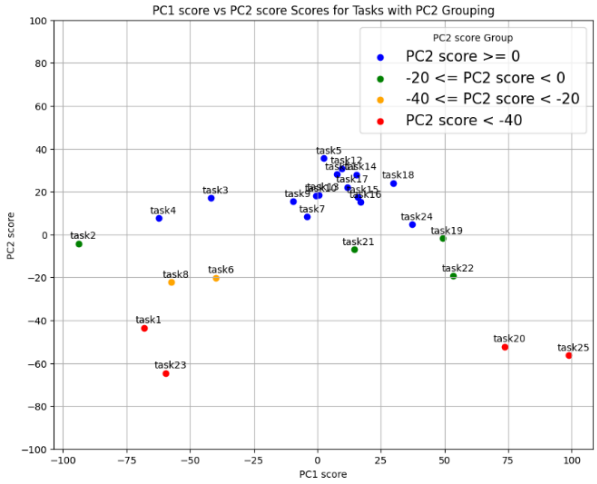


Figure 5: Task grouping based on PC2 scores.

Experiment 3: Reduction of Tasks and Assessment of Classification Performance

In this experiment, the input tasks were determined based on the task groups established in Experiment 2, and the measurements of the features associated with the selected tasks were used as explanatory variables. Each feature was standardized using the mean and standard deviation calculated from the training data, and the same transformation was applied to the validation and test data to prevent data leakage. Subject labels were used as the target variable, with individuals with AD coded as 1 and healthy individuals coded as 0. Three classification methods—Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—were employed. Accuracy, Sensitivity, and Specificity were used as evaluation metrics to assess changes in classification performance resulting from task reduction.

To account for fluctuations in classification performance caused by different training and test data splits, model training was performed five times using different data splits, and the average values were evaluated. The same data splits were used across all three models in each iteration. In addition, hyperparameters were optimized using a grid search with 5-fold cross-validation on the training data. The parameter ranges were determined based on prior research by Cilia et al., and the MLP was configured with three hidden layers. Table 1 lists the model parameters, and Tables 2–4 show Accuracy, Sensitivity, and Specificity.

Tables 2–4 show that Accuracy decreased for SVM and MLP as the number of tasks was reduced, whereas RF maintained relatively stable performance even with four tasks. Sensitivity also remained relatively high for RF after task reduction. The results of Experiment 3 suggest that the tasks identified by the proposed method as having small differences in motor characteristics between individuals with AD and healthy individuals had a relatively limited contribution to classification performance. In other words, as defined in this study, tasks exhibiting similar motor characteristics in both groups can be considered candidates for removal. Furthermore, these results suggest that quantifying task characteristics using the proposed method can support the selection of a reduced set of tasks while maintaining classification performance.

Table 1: Hyperparameter search ranges for RF, SVM, and MLP.

Model	Parameter	Candidate Values
RF	max_depth	3,4,5,6,7,8,9,10
	n_estimators	100,150,200,250,300
	bootstrap	True, False
	min_samples_split	2
	min_samples_leaf	1
SVM	kernel	rbf, linear
	C	0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5

(Continued)

Table 1: Continued.

Model	Parameter	Candidate Values
	gamma	0.5(RBF kernel only)
MLP	activation	relu, logistic, tanh
	hidden_layer_sizes	(20, 15, 10), (18, 12, 8), (16, 12, 8), (14, 10, 6)
	learning_rate_init	0.001, 0.005
	max_iter	1000
	tol	0.0001

Table 2: Accuracy of RF, SVM, and MLP (mean (standard deviation)).

Number of Tasks	RF	SVM	MLP
25	0.824 (0.058)	0.782 (0.066)	0.848 (0.080)
10	0.842 (0.081)	0.782 (0.097)	0.800 (0.082)
6	0.842 (0.101)	0.703 (0.045)	0.782 (0.058)
4	0.824 (0.066)	0.697 (0.077)	0.758 (0.096)

Table 3: Sensitivity of RF, SVM, and MLP (mean (standard deviation)).

Number of Tasks	RF	SVM	MLP
25	0.925 (0.055)	0.823 (0.090)	0.868 (0.102)
10	0.953 (0.055)	0.832 (0.108)	0.819 (0.110)
6	0.933 (0.046)	0.745 (0.087)	0.821 (0.074)
4	0.891 (0.030)	0.697 (0.093)	0.772 (0.094)

Table 4: Specificity of RF, SVM, and MLP (mean (standard deviation)).

Number of Tasks	RF	SVM	MLP
25	0.705 (0.072)	0.736 (0.055)	0.824 (0.067)
10	0.718 (0.124)	0.720 (0.139)	0.770 (0.108)
6	0.746 (0.155)	0.658 (0.070)	0.739 (0.058)
4	0.760 (0.135)	0.705 (0.164)	0.746 (0.119)

CONCLUSION

This study proposed a method for selecting handwriting tasks for removal based on their motor characteristics, with the aim of reducing the examination burden associated with Alzheimer's disease (AD) classification.

In the proposed method, measurements for each task-feature combination were represented as one-dimensional values using an Autoencoder (AE), and Principal Component Analysis (PCA) was applied to the resulting values. The results showed that the first principal component (PC1) reflected whether the measurement values of the task features tended to be higher for individuals

with AD or healthy individuals, whereas the second principal component (PC2) reflected the magnitude of the differences in measurement values between the two groups. Subsequently, tasks were progressively reduced based on their PC2 scores, and AD classification was performed using Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The results showed that RF maintained relatively stable classification performance after task reduction, whereas SVM and MLP showed some performance degradation as the number of tasks decreased. These findings suggest that the tasks identified as candidates for removal had a relatively limited contribution to classification performance and that task selection based on the distribution of motor characteristics may enable a reduction in the number of tasks while maintaining classification performance.

The proposed method is characterized by its ability to select tasks in advance based on differences in the distributions of motor characteristics across subjects, unlike conventional approaches that primarily select tasks based on classification performance. Therefore, the proposed method has the potential to identify the tasks required for classification before performing classification itself, without using classification performance as the criterion for task selection.

However, the actual burden experienced by participants was not directly assessed in this study. The reduction in examination burden was inferred from the reduction in the number of tasks. Therefore, future work should incorporate subjective assessments of task burden and objective indicators such as fatigue to determine whether task reduction using the proposed method leads to an actual reduction in participant burden.

REFERENCES

Alatrany, A.S., Khan, W., Hussain, A., Kolivand, H. and Al-Jumeily, D. (2024). "An explainable machine learning approach for Alzheimer's disease classification", Scientific Reports, Vol. 14, Article 2637. DOI: 10.1038/s41598-024-51985-w.

Babushkin, V., Alsuradi, H., Jamil, M.H., Al-Khalil, M.O. and Eid, M. (2023). "Assessing handwriting task difficulty levels through kinematic features: A deep-learning approach", Frontiers in Robotics and AI, Vol. 10, Article 1193388. DOI: 10.3389/frobt.2023.1193388.

Cilia, N.D., De Stefano, C., Fontanella, F. and Scotto Di Freca, A. (2018). "An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis", Procedia Computer Science, Vol. 141, pp. 466–471. DOI: 10.1016/j.procs.2018.10.141.

Cilia, N.D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A. and Parziale, A. (2022). "Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking", Engineering Applications of Artificial Intelligence, Vol. 111, Article 104822. DOI: 10.1016/j.engappai.2022.104822.

Gattulli, V., Impedovo, D., Pirlo, G. and Semeraro, G. (2023). "Handwriting task-selection based on the analysis of patterns in classification results on Alzheimer dataset", CEUR Workshop Proceedings, Vol. 3521.

Japanese Society of Neurology. (2017). Ninchishō shikkan shinryō gaidorain 2017 [Clinical Practice Guideline for Dementia 2017]. Website: https://www.neurology-jp.org/guidelinem/nintisyo_2017.html. Accessed 26 Jan 2026. (in Japanese).

Kavitha, C., Mani, V., Srividhya, S.R., Khalaf, O.I. and Tavera Romero, C.A. (2022). "Early-stage Alzheimer's disease prediction using machine learning models", Frontiers in Public Health, Vol. 10, Article 853294. DOI: 10.3389/fpubh.2022.853294.

Kindai University Hospital. (n.d.). Arutsuhaimā-byō, ninchishō no shindan, chiryō [Diagnosis and treatment of Alzheimer's disease and dementia]. Website: https://www.med.kindai.ac.jp/diseases/alzheimers_disease_dementia.html. Accessed 26 Jan 2026. (in Japanese).

Luria, G. and Rosenblum, S. (2012). "A computerized multidimensional measurement of mental workload via handwriting analysis", Behavior Research Methods, Vol. 44, No. 2, pp. 575–586. DOI: 10.3758/s13428-011-0159-8.

Moreira, A., Ferreira, A. and Leite, N. (2024). "Prediction of Alzheimer disease on the DARWIN dataset with dimensionality reduction and explainability techniques", in: Proceedings of the 1st International Conference on Explainable AI for Neural and Symbolic Methods, SciTePress, pp. 38–49. DOI: 10.5220/0013017400003886.

Slavin, M.J., Phillips, J.G., Bradshaw, J.L., Hall, K.A. and Presnell, I. (1999). "Consistency of handwriting movements in dementia of the Alzheimer's type: A comparison with Huntington's and Parkinson's diseases", Journal of the International Neuropsychological Society, Vol. 5, pp. 20–25.

World Health Organization. (2017). Global action plan on the public health response to dementia 2017–2025. Geneva: World Health Organization. Website: https://www.who.int/publications/i/item/9789241513487. Accessed 26 Jan 2026.