Predicting Alzheimer And Cognitive Impairment From Handwriting Tasks

Martijn van Dam, Jonathan Koop, and Merlin Urbanski
INFOMDA2

Abstract

The DARWIN handwriting dataset, comprising 174 participants and 450 handwriting-related features, provides a high-dimensional setting for predicting Alzheimer's disease. This report addresses the challenge of selecting features in such a setting, where traditional logistic regression methods fail due to the large number of predictors. We investigate whether principal component-based dimensionality reduction improves predictive performance compared to a naive significance-based feature selection approach. Specifically, we compare logistic regression models built on principal components (using both the elbow criterion and eigenvalue-greater-than-one rule) versus models built on the subset of features identified through independent t-tests. Our findings reveal that principal component approaches outperform the naive method in terms of accuracy and stability, while highlighting the importance of balancing model complexity and variance. We conclude that PCA-based dimensionality reduction is a more robust solution in high-dimensional settings, but care must be taken in deciding the optimal number of components.

1 Introduction

Neurodegenerative diseases, such as Alzheimer's, progressively impair brain and nerve function, affecting cognition and movement. This affects how people think, move, and perform on everyday tasks. There is no cure for such diseases, but early detection of these diseases could help slow the progression of symptoms and improve life for those affected (Cilia et al., 2022). One way to detect these diseases early on is by analyzing people's writing behavior (Cilia et al., 2022). Handwriting involves fine motor skills, suggesting that alterations in an individual's writing style could provide insight into their neurological health.

1.1 Data

In this report, we will draw on the DARWIN (Diagnosis of Alzheimer with Handwriting) data set (Fontanella, 2022), comprising handwriting and drawing samples from a total of 174 participants. Out of these, 89 were Alzheimer patients (AP), while the remaining 85 were labeled as healthy, thus serving as a control group. The mean age of the AP group was 71.5 years (SD = 9.50), and the mean age of the healthy group was 68.9 years (SD = 12.00). Participants completed 25 carefully designed tasks, ranging from simple geometric shapes (e.g., drawing circles or lines) to more complex activities such as copying sentences, drawing a house, and writing words from memory. These tasks were chosen to assess a range of motor and cognitive abilities. The participants performed these tasks with an ink pen on paper placed over a tablet capable of capturing high-resolution movement data, such as x-y coordinates, pressure, and timestamps, at a sampling rate of 200 Hz. This setup enabled the extraction of 18 features per task, which can be grouped into the following categories: temporal features (e.g., air time (pen not on paper)), kinematic features (e.g., mean speed of pen on paper), pressure features (e.g., average pen pressure on paper), tremor analysis (e.g., mean relative tremor for on-paper movements with pen), and spatial features (e.g., the number of continuous pen strokes on the paper). In total, 450 features (18×25) were collected across all tasks, each represented as continuous variables. Table 1 presents summary statistics of example features for the signature task. All tasks and features are described in the supplementary materials (Supplementary Materials A and B).

Table 1: Descriptive Statistics for Alzheimer and Healthy Groups for Five Features of the First Task (Signature Drawing)

	Alzheimer		Healthy	
	Mean	SD	Mean	SD
Air Time (ms)	6,507.02	13,351.85	4,781.65	11,893.90
Mean Speed on Paper (cm/s)	3.36	1.98	4.26	2.40
Mean Pressure (g)	1,594.82	322.30	1,765.82	189.90
Mean Relative Tremor	179.16	99.86	222.85	119.29
Number of Pen Downs	9.85	6.85	7.82	4.19

1.2 Research problem and question

It is evident that the data set has more dimensions than participants. Standard analysis for comparisons of groups or group membership (i.e., logistic regression analysis) with all features is, hence, not possible. However, one could select features that have been shown to be informative in discriminating AP from healthy subjects with mean differences between groups (Bagherzadeh-Khiabani et al., 2016). However, when comparing groups on their average feature values, the features are evaluated independently, leaving the feature's contribution to a model unknown. Moreover, repeatedly performing significance tests increases the risk of a type I error (Tabachnick and Fidell, 2013), rendering the described method unattractive. An alternative approach for handling high-dimensional data is Principal Component Analysis (PCA). PCA reduces dimensionality by identifying components that summarize the dimensions with the highest variance among all potential predictors (Bro and Smilde, 2014). This technique transforms the original features into a smaller set of uncorrelated variables, or principal components, which retain as much of the data's variance as possible. Dimension reduction methods such as PCA are particularly advantageous for simplifying complex datasets while preserving essential information (Ray et al., 2021).
In our assignment, we will evaluate the performance of logistic regression analysis with significance-based feature selection and principal components as predictors of the discrimination of AP from healthy subjects. The research question we aim to answer is:

Does logistic regression analysis with principal components improve predictive performance in classification compared to a classic logistic regression analysis with a selection of predictors?

In the Methods section of the report, we will explain the procedure of the analyses and elaborate on how we compare their outcomes. In the results section, the results are presented and discussed, accompanied by brief conclusions on the model performances. In the conclusion, the results are further discussed, and future research directions are described.

2 Methods

Building on these challenges of high dimensionality, we employed logistic regression to classify individuals as patients or healthy. Logistic regression was chosen due to its interpretability and suitability for binary classification tasks. However, the model requires the number of predictors to be smaller than the number of observations to allow the identifiability of the parameters. Given the high dimensionality of the data, three approaches for predictor selection were compared: a naive approach with feature selection based on independent t-tests and two principal component analysis (PCA)-based approaches for feature selection.

2.1 Benchmark: The Naive Approach

In the naive approach, predictors were selected based on statistical significance in independent t-tests comparing mean differences between patients and healthy individuals for each predictor, and variables with p-values below a specified threshold were included in the model.

At a standard significance level of $\alpha = 0.05$, 286 predictors were identified, surpassing the sample size and thus violating the condition of fewer predictors than observations. To address this, the significance threshold was systematically reduced to $\alpha = 0.001$, resulting in 168 selected predictors. Although this approach effectively reduced the number of predictors, it brings along notable limitations. Multiple testing increases the risk of Type I errors (Tabachnick and Fidell, 2013). Additionally, independent t-tests are problematic in the presence of collinearity, as they do not account for the shared variance among collinear predictors, potentially resulting in the selection of redundant variables (Tabachnick and Fidell, 2013).

2.2 Handling High Dimensionality With PCA

To overcome these limitations, PCA was employed as a dimensionality reduction technique. PCA transforms the original predictors into a smaller set of uncorrelated components that summarize the directions of greatest variance in the data by identifying linear combinations of the original features that maximize the variance in the data (James et al., 2013). The procedure begins by computing the covariance matrix of the predictors to quantify how each pair of features varies jointly. Eigenvalue decomposition is then applied to the covariance matrix, yielding eigenvectors (principal components) and eigenvalues, which represent the variance each component captures. The eigenvectors define the new axes onto which the data is projected, and the eigenvalues are used to prioritize these axes based on the variance they explain. PCA addresses multicollinearity, which we assume to be apparent in the data, since the same 18 features were measured for all 25 tasks. Additionally, since the components are ordered by the variance they explain, choosing a certain number of first components reduces the risk of overfitting by limiting the number of predictors while retaining key information. Standardization was performed prior to PCA to ensure that all predictors contributed equally. Two criteria were applied to determine the number of components to retain: the elbow method, identifying the point where additional explained variance plateaus, and the eigenvalue-greater-than-one rule, which retains components with eigenvalues exceeding one. Unlike the naive approach, PCA mitigates the risk of multiple testing and uses the data structure to identify patterns that may be obscured when evaluating predictors individually.

2.3 Evaluating Model Performance

The three approaches were evaluated using leave-one-out cross-validation (LOOCV). LOOCV is a prominently used method to assess how models generalize to unseen data. This is done by iteratively training the model on all observations except one and testing on the excluded instance. Unlike earlier analyses using the DARWIN data set that applied 5-fold cross-validation (Cilia et al., 2022), we opted for LOOCV to incorporate more observations in each model due to the limited sample size.

Predictive accuracy, defined as the proportion of correctly classified observations, was used as the primary performance metric. Additionally, Cohen's κ was calculated to account for agreement between predicted and actual classifications beyond chance. The model achieving the highest accuracy and κ was identified as optimal. To account for specificity

and sensitivity, we additionally visualized Receiver Operating Characteristic (ROC) curves and computed the Area Under Curve (AUC).

All analyses were carried out in R 4.4.2 (R Core Team, 2024) using the `tidyverse` (Wickham et al., 2019) and the `caret` (Kuhn and Max, 2008) package. The code is available under https://github.com/tinusvd/infomda2_assignment

3 Results

The benchmark logistic regression model with 168 predictors selected using a significance threshold of $\alpha = 0.001$, achieved an accuracy of 0.66 and a κ of 0.33. These findings suggest that the model's performance is just marginally superior to random guessing, which would result in an accuracy and κ of 0.5 and 0, respectively. This performance emphasizes the expected limitations of naive predictor selection approaches in high-dimensional settings. Additionally, the benchmark model frequently encountered convergence issues in the training process, likely caused by perfect separation due to the inclusion of a large number of predictors relative to the sample size. These issues call for the use of proper dimensionality reduction techniques.

To this end, principal component analysis conducted on all potential predictors to include in the model lead components explaining the variance shown in Figure 1.

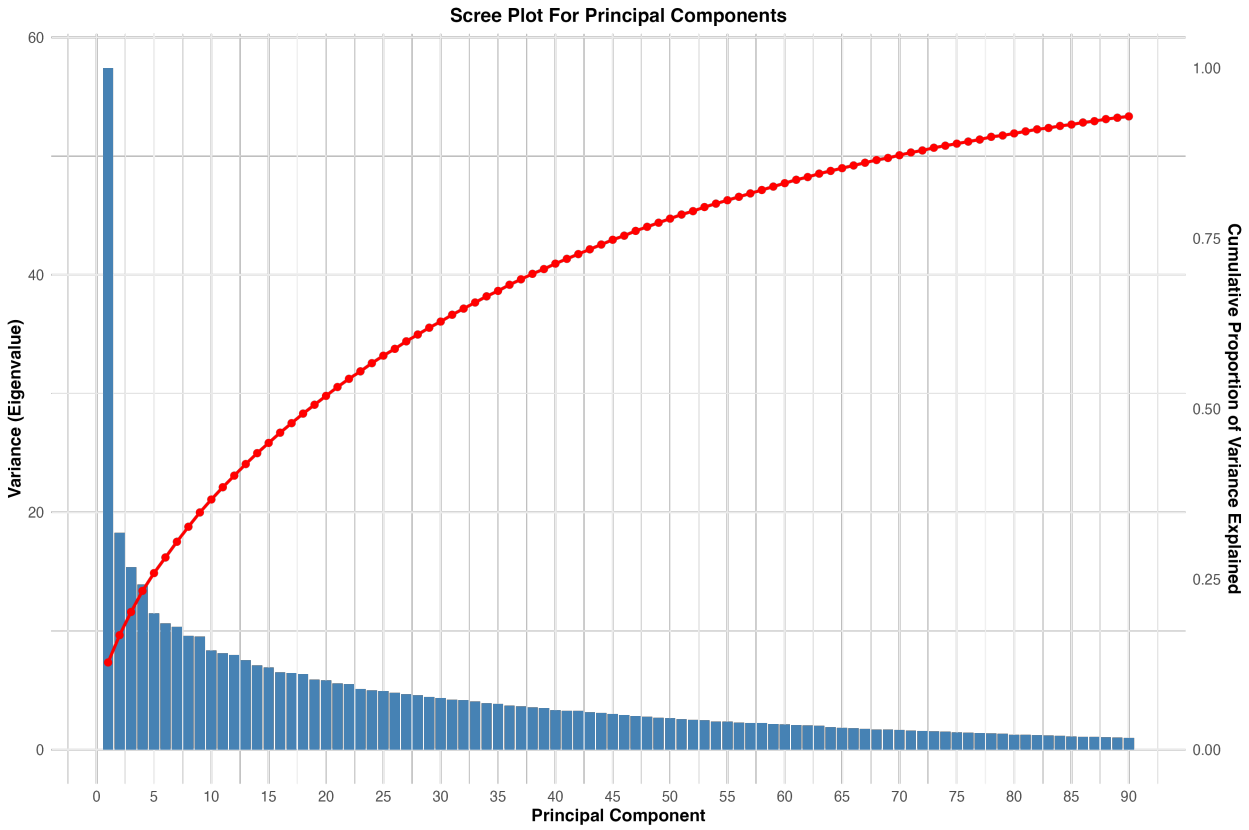

Figure 1: Scree plot displaying the eigenvalues of principal components and their cumulative explained variance. The horizontal axes show the component indices from 1 to 90. Vertical axes represent the eigenvalues (left axis) and cumulative proportion of variance explained (right axis). Blue bars correspond to the eigenvalues of individual components. The red line depicts the cumulative explained variance across components.

Based on the PCA-based elbow criterion, only the first principal component was selected as predictor. This decision was driven by the clear drop in eigenvalues for the second and subsequent components from 57.4 to 18.3. Although the elbow criterion typically suggests that further components add minimal explanatory power, the red line in the scree plot indicates that the cumulative explained variance still continues to increase beyond the first component. Notably, the first principal component alone accounts for only 12.8 % of the total variance. The logistic regression model trained on this component achieved an accuracy of 0.76 and a κ of 0.52, demonstrating moderate agreement between predictions and true values and a marked improvement over the benchmark model. By limiting the model to a single component, convergence issues were entirely avoided, demonstrating the utility of dimensionality reduction in reliable model estimation.

Using the eigenvalue-greater-than-one rule, 90 principal components were retained, collectively explaining 92.9% of the total variance. However, as visualized in Figure 1, the amount of explained variance that a component adds is very minimal when selecting many components. Nevertheless, the logistic regression model trained on these components achieved the highest accuracy of 0.82 and a κ of 0.64, indicating substantial agreement. Problematically,

this model encountered the same convergence issues as the benchmark model due to the inclusion of a large number of components. While retaining more components improved predictive performance, it also reduced numerical stability.

These results are further backed by AUC, which is highest for the model resulting from the eigenvalue-greater-than-one rule (0.90), second highest for that from the elbow criterion (0.86) and lowest for the naive approach (0.65). The respective ROC curves are shown in Figure 2.

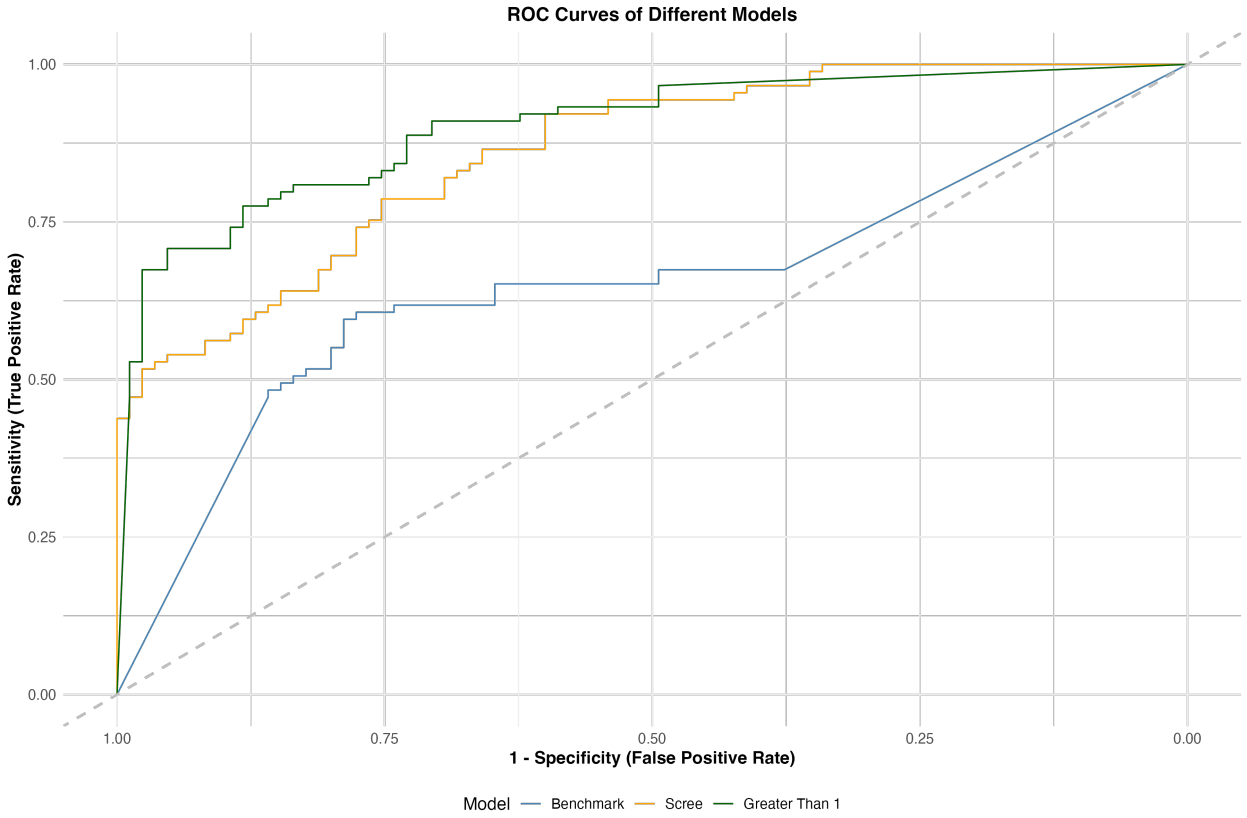

Figure 2: Receiver Operating Characteristic (ROC) curves for three logistic regression models. Horizontal axes show the False Positive Rate (FPR), while vertical axes represent the True Positive Rate (TPR). The dashed diagonal line corresponds to the performance of a random classifier (AUC = 0.5). The solid blue, orange, and green lines represent the ROC curves for models 1, 2, and 3, respectively. The Area Under the Curve (AUC) for each model quantifies the discriminatory ability, with higher values indicating better performance.

The two selection criteria effectively illustrate the bias-variance tradeoff: Retaining only one principal component, which explains 12.8% of the variance, results in an underfitting model that exhibits high bias due to insufficient information being captured. Conversely, retaining 90 components, which explain 92.9% of the variance, introduces substantial model complexity, increasing the risk of overfitting and high variance. Figure 3 illustrates this continuum, demonstrating the impact of iteratively adding components to the model on its accuracy. While accuracy does decline after about 50 components, surprisingly, it increases

once more after surpassing 80 components.

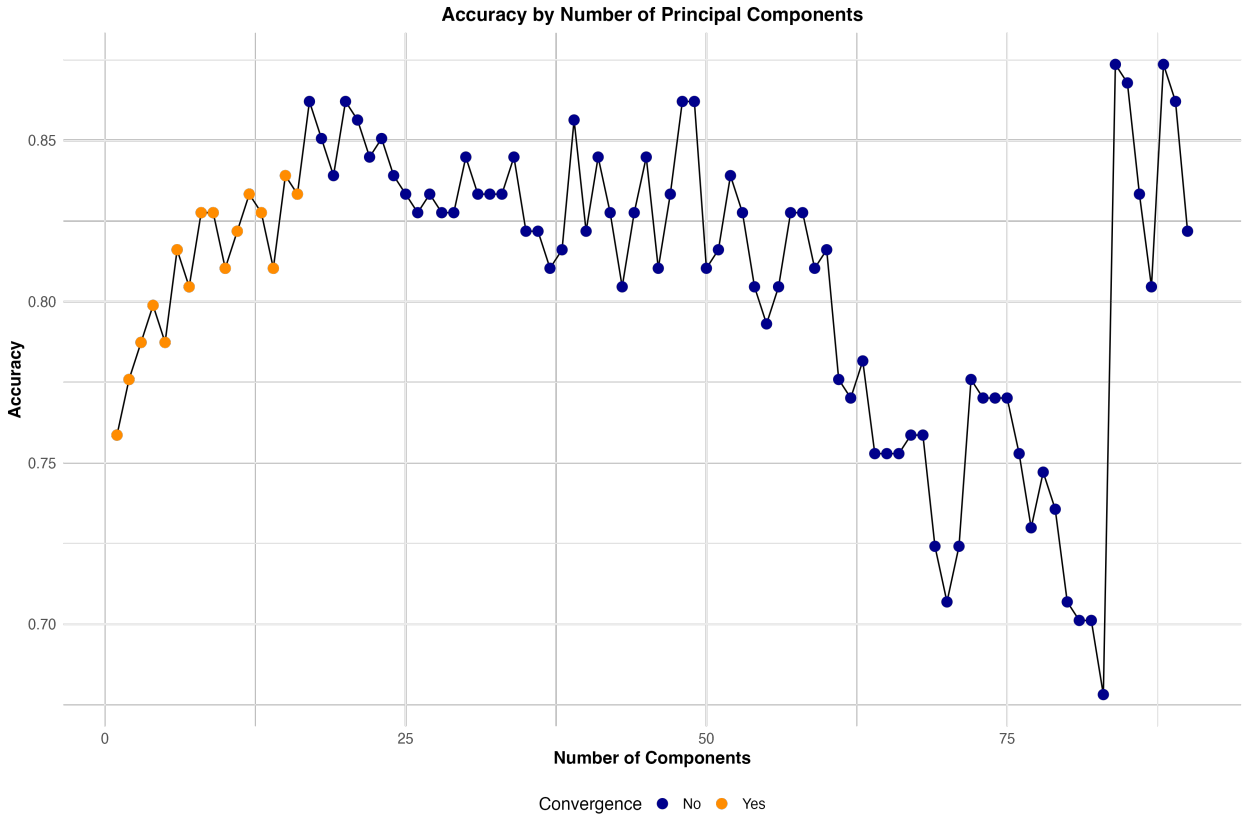

Figure 3: Scatter plot showing the accuracy by number of principal components used as predictors. Horizontal axes represent the number of principal components as predictors, while vertical axes indicate the accuracy of the respective model. Orange markers indicate that the model converged, and blue markers display that it did not.

4 Conclusion

This report aimed to evaluate the performance of high-dimensional modeling techniques in predicting Alzheimer's and cognitive impairment from handwriting tasks employing logistic regression models. To do so, principal component regression with component selection based on the scree plot and the eigenvalue-greater-than-1 rule was compared to predictor selection based on independent t-tests for the difference in means in the target variable. The resulting predictors were used to train logistic regression models, which were evaluated using accuracy and Cohen's κ from leave-one-out cross-validation. Regarding predictive performance, the two methods relying on principal component analysis demonstrated superior results to the naive approach, demonstrating its usefulness in highly dimensional settings.

Given that the elbow criterion leads to the selection of one principal component, while the eigenvalue-greater-than-1 suggested the selection of 90 components, the analysis resulted in models being on opposite sides of the bias-variance trade-off continuum: With the elbow criterion, an overly simplistic model is chosen, which leads to underfitting, and thus a high

model bias. In contrast, the model based on the eigenvalue-greater-than-1 criterion is overly complex with potentially higher variance. This implies future analyses should concentrate more on selecting fewer principal components for optimizing model complexity. As visualized in Figure 3 - the model performs similarly well with 20 to 50 principal components used as predictors compared to the more complex model resulting from the eigenvalue-greater-than-1 criterion. This is in line with expectations with concern to the bias-variance trade-off where models with balanced complexity perform best. The results additionally match with research done with regards to the optimal number of predictors used in logistic regression analyses: For instance, the prominently used one in ten rule (Harrell et al., 1984) would have led to the selection of fewer predictors given the relatively low sample size. As more contemporary literature has demonstrated this rule to not be reliable (van Smeden et al., 2016), the computation of other statistics may be necessary for identifying a more adequate number of predictors. Additionally, the unexpected increase in accuracy after 80 components should be analyzed in future research to uncover the meaning and cause.

Limiting the number of predictors would also help avoid convergence issues faced with the benchmark model and the principal component regression resulting from the eigenvalue-greater-than-1 criterion. As these issues are caused by fitted probabilities that are numerically 0 or 1 (also referred to as the *problem of separation*), approaches to tackling this problem may be viable in this context (Heinze and Schemper, 2002).

The ROC curve and corresponding AUC values provided insights into each model's performance in identifying Alzheimer's disease. The eigenvalue-greater-than-1 model achieved the highest AUC, indicating a strong discriminatory power in distinguishing Alzheimer's patients from healthy controls based on handwriting features. This implies that the model captures subtle changes in motor and cognitive functions, such as decreased pen control and tremor, which are early indicators of neurodegeneration. In contrast, the elbow criterion model, while achieving slightly lower AUC, also demonstrated meaningful discriminatory ability. The simpler model is easier to interpret, leveraging a single principal component to detect variations in handwriting that are associated with Alzheimer's disease. Because of its lower complexity, it is more practical for usage in clinical settings.

The naive approach, with its significantly lower AUC, highlights the limitations of independent feature selection methods in Alzheimer's diagnosis. Such methods do not account for the multicollinearity inherent in handwriting data, which reduces their potential to detect patterns that are indicative of early-stage neurodegeneration. Applied researchers should make an informed choice about the probability threshold for balancing sensitivity and specificity, depending on whether a missed diagnosis of Alzheimer's disease or an incorrect diagnosis of Alzheimer's disease is more important.

In light of the problems that the logistic regression models faced here, we also view the use of other models as a potential solution. In particular, since individuals were measured on 25 different tasks, the data here can be seen as nested. For this reason, the principal components could potentially be used as predictors in a multilevel logistic regression model where tasks are nested within individuals.

Altogether, this report thus illustrated the advantages of high-dimensional analysis techniques for handwriting data on the diagnosis of Alzheimer's, specifically those of principal component regression. However, in light of the limitations described, further research is needed.

References

Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., & Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology*, *71*, 76–85. https://doi.org/https://doi.org/10.1016/j.jclinepi.2015.10.002

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831. https://doi.org/10.1039/c3ay41907j

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

Fontanella, F. (2022). Darwin. https://doi.org/10.24432/C55D0K

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, *3*(2), 143–152. https://doi.org/10.1002/sim.4780030207

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409–2419. https://doi.org/10.1002/sim.1047

James, G. M., Witten, D., Hastie, T. J., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. Springer.

Kuhn & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: A review. *Artificial Intelligence Review*, *54*(5), 3473–3515. https://doi.org/10.1007/s10462-020-09928-0

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th). Pearson.

van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, *16*(1). https://doi.org/10.1186/s12874-016-0267-3

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

5 Supplementary Materials

A. Overview of Tasks

Participants in the DARWIN dataset completed 25 writing and drawing tasks. Below is a brief description of each task:

1. **Circle Retraction**: Drawing a circle multiple times to evaluate motor control.
2. **Signature Writing**: Writing their own signature to assess fine motor skills.
3. **Simple Line Drawing**: Drawing straight lines to measure basic motor precision.
4. **House Drawing**: Drawing a simple house to evaluate spatial and planning skills.
5. **Clock Drawing**: Drawing a clock to assess cognitive and spatial abilities.
6. **Word Copying**: Copying predefined words to measure visual-motor integration.
7. **Sentence Copying**: Copying sentences to test writing fluency and motor control.
8. **Free Word Writing**: Writing a word of their choice from memory.
9. **Number Copying**: Copying numbers to assess numeric literacy and motor precision.
10. **Number Writing (Dictation)**: Writing dictated numbers to test auditory-motor integration.
11. **Shape Copying**: Reproducing geometric shapes to measure spatial awareness.
12. **Name Writing**: Writing their own name to assess familiarity and motor control.
13. **Word Memory Writing**: Writing a word from memory to test recall and motor precision.
14. **Ellipsis Drawing**: Drawing repeated ellipses to analyze tremor and motor steadiness.
15. **Random Line Drawing**: Drawing random lines to evaluate exploratory motor skills.
16. **Spiral Drawing**: Drawing spirals to assess tremor and fine motor control.
17. **Pattern Copying**: Reproducing a predefined pattern to measure spatial and motor skills.
18. **Complex Figure Copying**: Copying a detailed figure to test integration of visual, motor, and cognitive functions.
19. **Sentence Writing (Dictation)**: Writing a dictated sentence to test auditory and motor integration.
20. **Word Recall (Timed)**: Writing as many recalled words as possible in a fixed time.
21. **Random Doodle**: Creating a random doodle to evaluate spontaneous motor skills.

22. **Repeated Letter Writing**: Writing a specific letter repeatedly to test consistency and motor precision.
23. **Drawing Shapes in Sequence**: Drawing shapes in a specific sequence to evaluate planning and motor coordination.
24. **Non-Dominant Hand Writing**: Writing with the non-dominant hand to measure motor asymmetry.
25. **Combined Task**: Performing a combination of drawing and writing tasks to measure multitasking abilities.

B. Overview of Features
For each task, the following 18 features were extracted:

- **Air Time**: Total time spent with the pen in the air.
- **Paper Time**: Total time spent with the pen on the paper.
- **Total Time**: Combined time spent on-paper and in-air.
- **Mean Speed (On-Paper)**: Average speed of the pen during on-paper movements.
- **Mean Speed (In-Air)**: Average speed of the pen during in-air movements.
- **Mean Acceleration (On-Paper)**: Average acceleration of the pen during on-paper movements.
- **Mean Acceleration (In-Air)**: Average acceleration of the pen during in-air movements.
- **Mean Jerk (On-Paper)**: Average rate of change of acceleration (jerk) during on-paper movements.
- **Mean Jerk (In-Air)**: Average jerk during in-air movements.
- **Pressure Mean**: Average pressure applied by the pen tip.
- **Pressure Variance**: Variability in the pressure applied by the pen tip.
- **GMRT (On-Paper)**: Generalized Mean Relative Tremor for on-paper movements.
- **GMRT (In-Air)**: Generalized Mean Relative Tremor for in-air movements.
- **Mean GMRT**: Combined mean of on-paper and in-air tremor metrics.
- **Pendown Count**: Number of distinct pen-to-paper movements.
- **X-Axis Extension**: Maximum horizontal movement distance.
- **Y-Axis Extension**: Maximum vertical movement distance.
- **Dispersion Index**: Measure of handwriting spread over the page.