Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine-Learning Modeling of DARWIN Handwriting Features for Alzheimer's Disease Classification: A Secondary Analysis

Sanjay Singhvi

Colridsanjay@gmail.com

University of Iowa https://orcid.org/0009-0005-2205-5453

Riddhi Singhvi

Stillwater Area High School

Research Article

Keywords: Alzheimer's disease, handwriting analysis, machine learning, feature selection, DARWIN dataset, secondary analysis

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10645598/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Handwriting dynamics have been investigated as a potential signal for Alzheimer's disease screening, but the standard DARWIN handwriting task set comprises 25 separate writing tasks. This secondary analysis asks a simple question: would a shorter set of tasks work just as well? Using the DARWIN dataset consisting of 174 participants (89 with Alzheimer's, 85 healthy controls), each performing 25 tasks that together yield 450 features, we found that a candidate subset of 10 tasks retained comparable performance in repeated internal evaluation. A Random Forest trained on this 10-task subset reached a test accuracy of 89.14% (95% CI: 87.33%–90.95%) and an AUC of 0.958. This result is close to the full 25-task set (88.95%). A paired Wilcoxon test did not detect a difference between the two ($p = 0.748$ across 30 repeated train/test splits). A non-significant result does not prove the two are equivalent. It does mean the shorter set performed comparably in this internal evaluation. Smaller subsets also showed no statistically detectable accuracy difference from the full 25-task set: 9 tasks ($p = 0.41$), 7 tasks ($p = 0.12$), and a 5-task consensus set ($p = 0.08$). As expected, deliberately poor subsets failed, and the 10 least-informative tasks lost 5.52 percentage points ($p = 0.0001$). We chose the tasks using three independent selection methods (RFE, LASSO, and mutual information). The three disagreed about which individual features mattered (Jaccard overlap of 0.06 to 0.36 at $k = 10$), but they agreed about which tasks mattered. All three ranked tasks 19, 23, 9, 25, 15, 24, 22, 17, 2, and 10 as the most informative, and the simple graphic and copying tasks (1 through 7) as the least. This pattern fits what is known about Alzheimer's: the signal concentrates in dictation and memory tasks, which draw heavily on working memory and executive function rather than on basic motor control. Within this cohort, the 10-task candidate set performed comparably to the full 25-task set on both accuracy and AUC. It uses 60% fewer tasks. A similar cut in testing time is likely but was not measured directly. Testing this on a separate, independent group of participants is the main remaining question and the goal of a planned follow-up study.

Introduction

Alzheimer's disease (AD) is the most common form of dementia. An estimated 6.5 million Americans aged 65 and older were living with it in 2022 (Alzheimer's Association, 2022; Tahami Monfared et al., 2022). Catching the disease early matters, because the treatments that can slow it work best in its earliest stages. The usual path to a diagnosis, however, has real limits: it takes clinician time, common screening tests such as the Mini-Mental State Examination miss many mild cases (Arevalo-Rodriguez et al., 2015), and brain imaging is expensive. There is strong interest, therefore, in screening tools that are cheap, non-invasive, and simple to administer.

Handwriting is one promising candidate. Writing is not a purely mechanical act; it draws at once on fine motor control, working memory, executive function, language, and visuospatial planning. Clinicians have noted for decades that people with dementia write in smaller letters (micrographia), move more slowly, and spend more time with the pen lifted off the page (De Stefano et al., 2019; Fernandes et al., 2023; Impedovo & Pirlo, 2019). Modern digitizing tablets record pen position, pressure, and timing many times

per second, which lets researchers measure the dynamics of writing, not just the finished marks on the page. A growing body of machine-learning work has used these dynamic measurements to separate people with AD from healthy controls, typically reaching accuracies between 75% and 95% on small, single-site groups.

The DARWIN dataset (Cilia et al., 2018, 2022; Fontanella, 2022) was created to fix two recurring problems in this field: samples that are too small, and acquisition protocols that differ from one study to the next. It contains 174 participants, with 89 with AD and 85 healthy controls matched for age and education, each completing the same 25 handwriting tasks on a Wacom Bamboo tablet. The tasks fall into four groups of rising cognitive difficulty: graphic tasks (drawing shapes, joining points), copying tasks (reproducing letters, words, and numbers), memory tasks (writing down previously memorized items), and dictation tasks (writing to spoken instructions, including a telephone number). Each task produces 18 measured features, so every participant is described by 450 features in total.

The paper that introduced DARWIN (Cilia et al., 2022) reported how well several standard models classified participants using all 450 features. Later work has extended it with explainable AI (Ho et al., 2025), stacking ensembles, and stroke-level analysis, and has repeatedly found that timing-related measurements, especially the time spent with the pen in the air and on the paper, are among the most informative.

Two questions about DARWIN have not yet been answered carefully. The first is how classification accuracy changes as the number of features grows, and whether there is a point past which extra features stop helping. If most of the useful signal sits in a small handful of tasks, a real-world screening test could be much shorter than the full 25-task version with no loss of accuracy, a practical win, since the time it takes to run the test is the main thing that limits whether clinics adopt it. The second question is whether different feature-selection methods agree on which features and tasks carry the signal. When independent methods agree, it is good evidence that a result is real rather than an accident of overfitting on a small dataset.

This study tackles both questions on the same data and under one consistent evaluation setup. It uses bootstrap confidence intervals, paired statistical tests, a saturating-curve fit to pin down where added features stop paying off, leave-one-out cross-validation as a check for the small sample, and a direct subset-validation experiment in which the candidate reduced task set is tested head-to-head against the full 25-task set.

Objectives

This study has four objectives:

- Measure how test accuracy and AUC change with the number of selected features (k) on DARWIN. We do this for three types of classifiers, namely Random Forest, L1-regularized Logistic Regression, and a small Multilayer Perceptron at eight values of k, with bootstrap 95% confidence intervals from

30 stratified 80/20 splits. We then identify the smallest k whose accuracy confidence interval overlaps that of the full feature set.

- Test whether three independent feature-selection methods, RFE with Logistic Regression, L1-Logistic Regression on all features, and mutual information with the class label, agree on which features and which tasks separate AD from controls.
- Locate the point of diminishing returns by fitting a saturating curve to each model's accuracy-versus-k relationship, and report the feature counts at which each model reaches half and 90% of its total gain.
- Evaluate a candidate reduced task set directly, by testing eight different task subsets on the same 30 splits and three classifiers, and checking with a paired Wilcoxon signed-rank test whether a 10-task subset drawn from the task ranking performs comparably to the full 25-task set.

Materials and Methods

Dataset

The DARWIN dataset (Fontanella, 2022) was obtained from the UCI Machine Learning Repository via Kaggle. It contains 174 participants in two groups: 89 labeled P (AD patient) and 85 labeled H (healthy control). Participants were recruited and assessed at the University of Cassino and Southern Lazio using standard cognitive tests: the MMSE, the Frontal Assessment Battery, and the Montreal Cognitive Assessment, and the two groups were matched for age and education (Cilia et al., 2022).

Each participant completed 25 standardized handwriting tasks on a Wacom Bamboo tablet sampling at 200 Hz. As above, the tasks cover four categories: graphic, copy, memory, and dictation. From each task, 18 features were extracted from the raw pen-stroke recordings. These include the time spent with the pen in the air, on the paper, and in total; the number of times the pen was lifted (pen-downs); the mean and maximum extent of writing along the x and y axes; mean speed and acceleration both in the air and on paper; mean jerk; mean Generalized Mean Relative Tremor (GMRT); and several pressure-based statistics. Stacking 18 features across 25 tasks gives 450 features per participant. The dataset has no missing values and no duplicate rows, and the two classes are close to balanced (51.1% vs 48.9%), so no resampling was needed.

Preprocessing

Two preprocessing steps were applied inside each training fold. First, each feature column was Winsorized at 5%: values below the 5th percentile or above the 95th percentile were capped at those percentile values (using scipy.stats.mstats.winsorize), which limits the influence of extreme outliers. Second, all features were standardized to zero mean and unit variance with scikit-learn's StandardScaler. Both steps were fit on the training data only and then applied to the test data, so no test information leaked into training.

Feature-Selection Methods

Three feature-selection methods were used. Each was fit on the training fold only, used to pick a top-k set of features, and then evaluated on the test fold.

- RFE (Recursive Feature Elimination) with Logistic Regression as the base model (L2 penalty, C = 1, liblinear solver, max_iter = 2000), removing 10 features per round until k remained.
- LASSO: an L1-regularized Logistic Regression (L1 penalty, C = 1, liblinear, max_iter = 2000) trained on all 450 features, keeping the top k by absolute coefficient size.
- MI: SelectKBest with mutual_info_classif, keeping the top k features by mutual information with the class label.

Classifiers

Three types of classifiers were trained and evaluated at every value of k, all implemented in scikit-learn (Pedregosa et al., 2011).

- Random Forest (RF): 300 trees, no depth limit, min_samples_leaf = 1, class_weight = balanced.
- L1-regularized Logistic Regression (LR-L1): L1 penalty, C = 1, liblinear, max_iter = 2000, class_weight = balanced. This model is a separate step from the LASSO selector above; the classifier is trained on the features that were already selected.
- Multilayer Perceptron (MLP): one hidden layer of 32 ReLU units, L2 regularization (alpha = 1e-3), early stopping on a 15% internal validation split with patience 20, max_iter = 500.

Train/Test Split and Evaluation

For each of 30 random seeds, a single stratified 80/20 train/test split was drawn (the seed set the random_state). With 174 participants, that gives about 139 training and 35 test participants per seed. Feature selection, scaling, and classifier training all happened inside the training portion only. Crucially, every model, selector, and task subset used the same train/test boundary for a given seed. Because they share that boundary, all comparisons are paired by seed and can be analyzed with a paired statistical test.

Six metrics were recorded for each model, condition, and seed: test accuracy, recall for the positive (P) class, precision (P), macro F1, AUC, and Brier score. Each condition is summarized as the mean over 30 seeds, with bootstrap 95% confidence intervals from 2000 resamples. Models and subsets were compared with paired Wilcoxon signed-rank tests on the per-seed accuracy differences. As an extra check for the small sample, the Random Forest at k = 10 was also evaluated by leave-one-out cross-validation, with feature selection repeated inside each fold.

Saturating-Curve Fit

To find where extra features stop helping, we fit each model's mean accuracy as a function of k with a Hill function: $y(k) = y_{min} + (y_{max} - y_{min}) \cdot k / (K_{half} + k)$, using non-linear least squares (scipy.optimize.curve_fit). The half-saturation point, $K_{50} = K_{half}$, is the number of features at which half of the total accuracy gain has been reached. The 80% and 90% points follow as $K_{80} = 4 \cdot K_{half}$ and $K_{90} = 9 \cdot K_{half}$.

Task-Subset Validation

To test directly whether a focused set of tasks can match the full 25-task set, we defined eight task subsets and evaluated them under the same 30-seed setup as the main analysis.

The reference condition is ALL25_FULL: all 25 tasks and all 450 features, with no further feature selection. The main candidate reduced task set is TOP10_MEAN: the 10 tasks ranked highest by average importance across the three selection methods (tasks 19, 23, 9, 25, 15, 24, 22, 17, 2, and 10), giving 10 × 18 = 180 features.

Six other conditions were added to map out the landscape: TOP7_MEAN (top 7 tasks, 126 features); MAJORITY9 (the 9 tasks that appear in the top 10 of at least 2 of the 3 methods, 162 features); CONSENSUS5 (the 5 tasks in the top 10 of all 3 methods: 9, 15, 19, 23, 25, with 90 features); ALL25_K25_RFE (all 25 tasks but only 25 RFE-selected features); ORIGINAL_T1_T25 (just tasks 1 and 25, the first and last in the protocol, 36 features, as a minimal baseline); and BOTTOM10 (the 10 lowest-importance tasks, 180 features, as a sanity check). Each subset used the same 30-seed splitting, scaling, and three-classifier evaluation as the main analysis, and was compared with paired Wilcoxon signed-rank tests against both ALL25_FULL and ALL25_K25_RFE.

Results

Feature-Count Curve

Figure 1 shows test accuracy and AUC as a function of the number of selected features, k, for all three classifiers, with bootstrap 95% confidence bands. Random Forest is the best classifier across the whole range. Its accuracy climbs from 77.90% [75.43%, 80.38%] at k = 3 to 88.95% [87.05%, 90.57%] at k = 450, a span of 11.05 percentage points, with AUC rising from 0.857 to 0.963. The curve is strongly concave: most of the gain happens at low k, and then it flattens out. Using the rule that k is "enough" once its 95% confidence interval overlaps the k = 450 interval for all three classifiers at once, k = 25 is the smallest sufficient feature count from per-feature selection.

Figure 1

Test Accuracy and AUC as a Function of the Number of Selected Features

Note

Test accuracy (left) and AUC (right) versus the number of features, k, for Random Forest (circles, solid blue), L1-regularized Logistic Regression (squares, dashed black), and a small Multilayer Perceptron (triangles, dotted gray). Lines are the mean over 30 stratified 80/20 splits; shaded bands are bootstrap 95% confidence intervals. The x-axis is on a log scale. Different markers, line styles, and colors keep the three models distinguishable in grayscale. Random Forest leads throughout, and the accuracy curve is strongly concave: at k = 25 the confidence intervals overlap the k = 450 intervals for all three models, but at k = 10 they do not.

Saturating-Curve Fit

Table 1 gives the Hill saturation fit for each model. For Random Forest, half the accuracy gain arrives by just 4 features (K_{50} = 4.0; residual RMSE 0.002), 80% by 16 features, and 90% by 36. The linear model saturates more slowly (LR-L1: K_{50} = 11.9, K_{90} = 107), and the MLP slower still (K_{50} = 8.9, K_{90} = 80). In short, Random Forest extracts the signal more efficiently than the linear model or the small neural network at any given number of features.

Table 1
Hill Saturation Fit to Mean Test Accuracy

Model	y_{min}	y_{max}	K_{50}	K_{80}	K_{90}	RMSE
RF	0.697	0.889	4.0	16	36	0.002
LR-L1	0.749	0.879	11.9	48	107	0.005
MLP	0.611	0.822	8.9	35	80	0.013

Note. Hill saturation fit to mean test accuracy as a function of k: y(k) = ymin + (ymax − ymin) · k / (Khalf + k). K50 = Khalf is the number of features at which half of the total accuracy gain is reached. Random Forest saturates fastest of the three classifiers.

Per-Feature Selection: Headline at k = 25

Table 2 reports each classifier's mean performance and bootstrap 95% confidence intervals over 30 seeds at k = 25, the smallest feature count from per-feature selection that meets the overlap rule above. Random Forest is best on every metric: accuracy 86.48% [84.19%, 88.67%], AUC 0.934 [0.918, 0.949], macro F1 0.864, recall (P) 0.880, precision (P) 0.868, and Brier score 0.109. Paired Wilcoxon tests confirm that all model differences at k = 25 are significant: RF beats LR-L1 by 2.48 percentage points on average (p = 0.013), RF beats MLP by 8.10 points (p < 0.001), and LR-L1 beats MLP by 5.62 points (p = 0.002). The ordering RF > LR-L1 > MLP holds at k = 25.

Table 2

Per-Classifier Test Performance at k = 25

Model	Accuracy	AUC	Macro F1	Recall (P)	Brier
RF	0.865 [0.842, 0.887]	0.934 [0.918, 0.949]	0.864 [0.841, 0.886]	0.880 [0.850, 0.907]	0.109 [0.100, 0.119]
LR-L1	0.840 [0.817, 0.862]	0.901 [0.884, 0.918]	0.839 [0.816, 0.861]	0.811 [0.782, 0.839]	0.126 [0.110, 0.143]
MLP	0.784 [0.749, 0.815]	0.863 [0.831, 0.890]	0.781 [0.746, 0.813]	0.781 [0.741, 0.822]	0.153 [0.139, 0.168]

Note. Per-classifier test performance at k = 25 (RFE-selected features). Values are the mean over 30 stratified 80/20 splits, with bootstrap 95% confidence intervals in brackets. The positive class is P (AD patient). N(train) = 139, N(test) = 35 per seed.

Feature-Count Effect Compared to Model Effect

How much does the number of features matter compared with the choice of model? Increasing k from 3 to 450 raises accuracy by 11.05 points for RF, 9.71 for LR-L1, and 12.76 for MLP. Switching between the best and worst model at a fixed k has a similar-sized effect: a gap of 10.29 points at k = 3, peaking at 13.05 points at k = 6, narrowing to 5.33 at k = 100, and widening again to 8.57 at k = 450.

The two effects are of comparable size. At every k tested, deciding how many features to use matters about as much as deciding which model family to use. Once feature selection has done its job (k ≥ 50), the three models sit within roughly 9 points of each other. The practical lesson is that getting the feature set right is at least as important as picking the model. For clinical screening, where interpretability counts, this argues for spending effort on feature selection and task-set design rather than on model architecture.

Selection-Method Agreement

Figure 2 shows the Jaccard overlap between the top-k feature sets chosen by each pair of methods, averaged over 30 seeds, at k = 6, 10, and 25. RFE and LASSO, which both use a Logistic Regression base, agree moderately (Jaccard 0.28 at k = 6, 0.36 at k = 10, 0.53 at k = 25). Mutual information disagrees with both (Jaccard 0.06 to 0.10). At k = 10, only two features appear in the top-10 lists of all three methods: total_time15 and air_time22.

Figure 2

Agreement Between Feature-Selection Methods

Note

Jaccard overlap between the top-k feature sets chosen by RFE, LASSO, and mutual information (MI) at k = 6, 10, and 25, averaged over 30 seeds with standard-deviation error bars. RFE and LASSO, both Logistic-Regression-based, agree moderately. MI disagrees with both at the feature level.

This low feature-level agreement does not mean any one method is wrong. RFE and LASSO pick features in context; each candidate is judged alongside the features already chosen, while mutual information scores each feature on its own. When useful features are correlated, the context-aware methods keep one and drop its correlates, whereas mutual information may keep several correlated features that share the same signal. So the three methods are really answering slightly different questions. As the next section shows, they nonetheless agree closely on which tasks matter.

Task-Level Importance

Every one of the 450 features belongs to one of the 25 tasks. When per-feature importance is rolled back up to the task level (Fig. 3), the three methods agree strongly. Task 19 is the single most important task under all three, accounting for 9.7% of RFE-selected features, 12.0% of LASSO-selected features, and 25.9% of MI-selected features in the top 25. Tasks 23 (writing a telephone number to dictation), 9, 25, 15, 24, and 22 are also among the most informative under all three methods, each carrying between 4% and 10% of the importance. Ranking tasks by their mean importance across methods gives this top 10: 19, 23, 9, 25, 15, 24, 22, 17, 2, 10. At the other end, tasks 1 through 7, the simple graphic and copying tasks at the start of the protocol, along with task 18, are consistently the least informative under all three methods.

Figure 3

Task-Level Importance Under Three Feature-Selection Methods

Note

Task-level importance for each of the 25 DARWIN tasks under the three feature-selection methods. For each method, the 18 features per task were summed and normalized within that method, then averaged over 30 seeds at k = 25. Despite low feature-level agreement (Fig. 2), all three methods rank task 19 as most important and tasks 23, 9, 25, 15, 24, and 22 as jointly informative, while tasks 1 through 7 are consistently low.

Permutation Importance and Feature Stability

Permutation importance was computed on the test fold for the Random Forest at k = 10 across all 30 seeds. The most consistently selected and most permutation-important features were total_time15, air_time22, and num_of_pendown19 (each chosen in 25 of 30 seeds), followed by total_time9, total_time2, and pressure_var19 (chosen in 10 to 16 of 30 seeds). All of these are timing- or count-related, in line with the wider DARWIN literature that timing variables dominate the AD signal (Cilia et al., 2022; Ho et al., 2025). Notably, none of them come from tasks 1 through 7.

Calibration

Figure 4 shows the reliability diagram pooled across 30 seeds at k = 25, with 1050 held-out predictions per model. Random Forest tracks the diagonal, the line of perfect calibration, closely across most of the range. The L1 Logistic Regression is somewhat overconfident at the extremes, and the MLP stays close to the diagonal throughout. Mean Brier scores at k = 25 are 0.109 (RF), 0.126 (LR-L1), and 0.153 (MLP). These pooled predictions come from the same 174 participants across repeated splits, so the diagram should be read as a rough guide only. Calibration for any clinical use would need to be checked on independent, out-of-sample predictions first.

Figure 4

Reliability Diagram for the Three Classifiers at k = 25

Note

Reliability diagram for the three classifiers at k = 25. Predicted probabilities were sorted into 10 quantile-based bins, and the observed fraction of positives is plotted against the mean predicted probability in each bin. Pooled across 30 seeds (1050 predictions per model). The diagonal marks perfect calibration.

Leave-One-Out Cross-Validation Sanity Check

As a check for the small sample, the Random Forest at k = 10 was re-evaluated with leave-one-out cross-validation, repeating feature selection inside each of the 174 folds. This run gave accuracy 83.33%, recall (P) 85.39%, precision (P) 82.61%, AUC 0.922, and Brier 0.116. The matching 30-seed 80/20 estimate at k = 10 was accuracy 83.43%, recall 85.37%, and AUC 0.911. The two approaches agree to within 0.1 points on accuracy and 0.01 on AUC. This close agreement provides a useful robustness check against unusually favorable or unfavorable individual test splits. This analysis concerns the k = 10 per-feature model and is separate from the 10-task TOP10_MEAN analysis.

Task-Subset Validation

Figure 5 and Table 3 present the main practical result. The candidate 10-task set (TOP10_MEAN: tasks 19, 23, 9, 25, 15, 24, 22, 17, 2, 10; 180 features; no further feature selection) reaches a Random Forest test accuracy of 89.14% [87.33%, 90.95%] and an AUC of 0.958. The full 25-task set reaches 88.95% [87.05%, 90.57%] and an AUC of 0.963. The paired Wilcoxon test on per-seed accuracy differences gives W = 96.5, p = 0.748 (n = 30). In other words, the test found no detectable accuracy loss for the 10-task set in this repeated internal evaluation. This absence of detected difference is not a formal proof of equivalence. Against ALL25_K25_RFE, the same 10-task set is 2.67 points better on accuracy (W = 72.0, p = 0.014), suggesting that task-level selection may outperform aggressive individual-feature selection in this dataset.

Figure 5

Test Accuracy and AUC for Each Task Subset

Note

Random Forest, L1-Logistic Regression, and MLP test accuracy (left) and AUC (right) for each of the eight task subsets, shown as the mean over 30 stratified 80/20 splits with bootstrap 95% confidence intervals as error bars. ALL25_FULL is the reference. TOP10_MEAN, MAJORITY9, TOP7_MEAN, and CONSENSUS5 all reach RF accuracy within sampling noise of the reference. ORIGINAL_T1_T25 (the first and last tasks) and BOTTOM10 (the 10 least-informative tasks) are clearly worse, confirming the task ranking. Different markers, fill colors, and hatch patterns keep the three models readable in grayscale.

The result holds at every further level of compression. MAJORITY9 (9 tasks, 162 features) reaches 88.00% (0.95 points below full, $p = 0.407$). TOP7_MEAN (7 tasks, 126 features) reaches 87.14% (1.81 points below, $p = 0.124$). CONSENSUS5 (5 tasks, 90 features) reaches 87.24% (1.71 points below, $p = 0.082$). None of these differ significantly from the full 25-task set.

The deliberately weak subsets behave as expected. BOTTOM10 (the 10 least-informative tasks, 180 features) reaches only 83.43%, 5.52 points below full ($p = 0.0001$), providing supporting evidence that the task ranking captures informative differences within this cohort. ORIGINAL_T1_T25 (just the first and last tasks, 36 features) reaches only 75.43%, 13.52 points below full ($p < 0.0001$) and worse even than ALL25_K25_RFE.

One result runs against standard feature-selection intuition: TOP10_MEAN (180 features fed straight to RF) beats ALL25_K25_RFE (25 carefully selected features from the full 450) by 2.67 points. When participants are few relative to the number of candidate features, giving a Random Forest a larger pool of features performed better in this analysis than aggressive per-feature selection, as long as that larger pool is drawn from the informative tasks. Two explanations fit, and both may hold. First, the per-feature selectors are fit on only 139 training participants and may over-select correlated features; restricting to whole tasks acts as a coarser, more stable form of regularization. Second, Random Forest's bagging-with-feature-subsampling works better with 180 features than with 25, because the extra features make individual trees less correlated with each other.

Table 3

Random Forest Test Performance by Task Subset

Subset	Tasks	# Features	RF accuracy [95% CI]	RF AUC	p (vs FULL)
ALL25_FULL	25	450	0.890 [0.871, 0.906]	0.963	reference
TOP10_MEAN	10	180	0.891 [0.873, 0.910]	0.958	0.748
MAJORITY9	9	162	0.880 [0.857, 0.901]	0.964	0.407
CONSENSUS5	5	90	0.872 [0.853, 0.891]	0.945	0.082
TOP7_MEAN	7	126	0.871 [0.851, 0.892]	0.946	0.124
ALL25_K25_RFE	25	25 (RFE)	0.865 [0.842, 0.887]	0.934	0.011
BOTTOM10	10	180	0.834 [0.811, 0.856]	0.921	0.0001
ORIGINAL_T1_T25	2	36	0.754 [0.732, 0.777]	0.847	< 0.0001

Note. Random Forest test performance by task subset, as the mean over 30 stratified 80/20 splits with bootstrap 95% CIs in brackets. The rightmost column gives the paired Wilcoxon signed-rank p-value against ALL25_FULL on per-seed accuracy differences (n = 30 paired splits). Candidate reduced subsets are listed first, followed by the feature-selection baseline and deliberately weak comparison subsets.

Discussion and Conclusions

Four findings stand out.

First, a 10-task subset drawn from the task ranking performed comparably to the full 25-task set. The paired test detected no accuracy difference between the 10-task and 25-task sets ($p = 0.748$); their AUC values were also descriptively similar. This pattern held all the way down to 5 tasks (CONSENSUS5: $p = 0.082$). It only broke when the tasks were drawn from the least-informative end (BOTTOM10: $p = 0.0001$) or cut to a bare 2-task baseline (ORIGINAL_T1_T25: $p < 0.0001$). The result comes from internal evaluation on one DARWIN cohort, not external validation on a new one. That distinction matters for how strongly the claim can be stated (see Limitations). It is also important that a non-significant test does not prove equivalence. Within this cohort, the finding is best read as preliminary: a focused 10-task set performed comparably to the full 25-task set while using 60% fewer tasks.

Second, the link between the number of features and Random Forest accuracy is strongly concave. The Hill fit puts half-saturation at 4 features, 80% saturation at 16, and 90% at 36, with $k = 25$ the smallest

count whose confidence interval overlaps the full feature set for all three classifiers. The 10-task result fits the same theme at a coarser level: most of the signal lives in a small part of the DARWIN task set, whether measured feature-by-feature or task-by-task.

Third, the three feature-selection methods, two that judge features in context and one that judges them in isolation, disagree at the feature level (Jaccard 0.06 to 0.36 at k = 10) but agree strongly at the task level. Task 19 is the top task under all three, and tasks 23, 9, 25, 15, 24, and 22 are jointly informative under all three. Agreement among methods provides supporting evidence that the task-level pattern is not specific to a single selection technique, although it does not eliminate the possibility of cohort-specific overfitting.

Fourth, the ranking of tasks lines up with what is known about Alzheimer's. Tasks 1 through 7 are simple graphic tasks (joining points, drawing shapes) that mainly test basic motor control and hand-eye coordination. Tasks 8 through 14 are copying tasks of rising difficulty. Tasks 15 through 25 include memory and dictation tasks that progressively lean harder on working memory and executive function; task 23, for instance, is writing a telephone number to dictation, and tasks 19, 22, 24, and 25 fall in similar memory-and-dictation categories. The selection methods converge on this later part of the protocol. Alzheimer's does affect fine motor control, but in this cohort the strongest early signal lies along the cognitive-demand axis: working memory, executive function, and planning complex movements under mental load (De Stefano et al., 2019; Fernandes et al., 2023). The finding that simple graphic tasks carry little signal while demanding dictation tasks carry most of it fits this cognitive picture rather than a pure motor-impairment one. The dominance of timing features (air time, paper time, total time, pen-down counts) within the informative tasks reinforces this: on cognitively demanding tasks, people with AD are slower and hesitate more, and timing features capture exactly that hesitation.

Finally, TOP10_MEAN (10 tasks, 180 features fed directly to RF) outperformed ALL25_K25_RFE (all 25 tasks but only 25 selected features) by 2.67 points ($p = 0.014$). When the sample is small relative to the number of candidate features, feeding a Random Forest a moderately large pool of features drawn from the informative tasks appeared to work better than aggressive per-feature selection in this dataset. This advantage may be because per-feature selectors fit on 139 participants over-select correlated features, while restricting to whole tasks is a coarser and steadier form of regularization; and because Random Forest's bagging benefits from a larger feature pool through less-correlated trees. The broader takeaway for small biomedical datasets like this one is worth stating plainly: it can be worth trying domain-informed, coarse selection before reaching for fine-grained statistical selection.

Limitations

Several limitations apply.

- **Internal validation only.** The 10-task subset and the full 25-task set were both tested on the same DARWIN cohort. The result ($p = 0.748$) means we found no accuracy loss for the subset within this cohort. It does not prove the two are equivalent, and it does not show that the subset will hold up on

a new group of participants recorded on different equipment. External validation on an independent cohort is the main remaining question and the focus of a planned follow-up study.

- **Task selection reused the same data.** The task ranking was built from mean feature importance across all 30 splits, and the resulting 10-task set was then scored on those same splits. Feature selection was redone inside each training fold, but the fixed task ranking still drew on the whole cohort. Because of this reuse, the reported 10-task accuracy may be optimistic. A nested design — deriving the task ranking inside inner folds and scoring once on an untouched outer fold — would give a cleaner estimate and is planned as future work.
- **The 30 splits are not independent.** The 30 repeated train/test splits share most of their participants. They are not 30 separate experiments. Treating them as independent in the Wilcoxon tests and bootstrap intervals can make the uncertainty look smaller than it is. The confidence intervals and p-values should be read with this in mind.
- **Single site, single protocol.** All DARWIN participants were recruited at one center in Italy, using one Wacom Bamboo tablet and one acquisition protocol. Whether the results carry over to other languages, writing systems, tablet hardware, and patient populations has not been tested.
- **Class definition.** The patient class is “clinically diagnosed with AD”; it does not separate disease stages or subtypes, and Mild Cognitive Impairment is not included as its own class. The healthy class is defined as the absence of a diagnosis. The binary AD-versus-control accuracies reported here should not be read as estimates for the harder task of separating MCI from healthy controls.
- **Small sample.** DARWIN has 174 participants. Under an 80/20 split, each test fold holds about 35 people, so a single misclassification moves test accuracy by about 2.9 points. The leave-one-out agreement reported above shows the results are not just split-size noise, but caution is still warranted when comparing accuracies that differ by less than three or four points.
- **The selection methods do not cover everything.** RFE-LR, LASSO, and MI span both multivariate-linear and univariate-information-theoretic approaches. Still, they leave out tree-based selection (Random Forest mean decrease in impurity, with corrections for high-cardinality bias) and stability selection with bootstrap subsampling. Adding a tree-based selector could shift the task-level agreement.
- **The MLP was not heavily tuned.** It used one fixed architecture (a single 32-unit hidden layer, alpha = 1e-3, early stopping). A wider search over layer sizes, regularization, and learning rates might raise its performance, though given the small sample the room for improvement is probably modest.
- **Task time was not measured directly.** The subset uses 60% fewer tasks, but this is not the same as a 60% cut in testing time. The 25 tasks may not take equal time to administer. The original protocol paper (Cilia et al., 2018) does not report per-task durations in a form that can be added up cleanly. Because the TOP10_MEAN subset mixes task categories, the count-based figure is at least not skewed toward any one category. The safe reading is that the subset uses 60% fewer tasks; the actual time saving is likely but was not measured, and should be recorded during the planned validation study.

Future Work

These limitations point to several next steps.

- **External validation.** Test the 10-task TOP10_MEAN subset prospectively on a new cohort with matched demographics, the same Wacom Bamboo protocol, and the same feature-extraction pipeline. The planned prediction is that the new cohort's Random Forest accuracy and AUC on the 10-task subset will fall within the current bootstrap 95% intervals (accuracy 87.33%–90.95%, AUC 0.945–0.970). External validation is the main open question following from the present result.
- **MCI extension.** Add a Mild Cognitive Impairment class as an intermediate target, to test whether the same task ranking holds for the harder job of separating MCI from healthy controls.
- **Tree-based selection.** Add mean-decrease-in-impurity selection from a Random Forest, or stability selection with bootstrap subsampling, to check whether the task-level agreement survives a different family of selectors.
- **Cross-cohort generalization.** Apply the trained TOP10_MEAN pipeline to a second AD-handwriting cohort (Chen et al., 2024) to test whether it generalizes across cohorts, the key open question in this field.
- **Direct timing measurement.** Record per-task acquisition times on a new cohort, to turn the 60% cut in task count into a precise time-saving figure.
- **Better calibration.** Although Random Forest at k = 25 is already reasonably calibrated, applying isotonic regression or Platt scaling on a held-out fold would tighten the high-probability tail before any clinical use.
- **Multimodal extension.** Combine DARWIN-style handwriting features with structural MRI features from a paired cohort, or with cognitive scores such as the MMSE, to test whether handwriting adds information beyond the standard diagnostic workup.

Summary

This study reported a feature-efficiency analysis of the DARWIN handwriting dataset, using three independent feature-selection methods, three classifier families, eight feature-set sizes, 30 seeds of stratified splitting, bootstrap 95% confidence intervals, paired Wilcoxon signed-rank tests, a Hill saturation fit to quantify diminishing returns, leave-one-out cross-validation as a small-sample check, and a direct subset-validation experiment on eight task subsets.

The headline finding is that a 10-task subset of DARWIN, tasks 19, 23, 9, 25, 15, 24, 22, 17, 2, and 10, giving 180 features fed to a Random Forest with no further feature selection, reaches test accuracy 89.14% [87.33%, 90.95%] and AUC 0.958. A paired Wilcoxon test found no detectable difference from the full 25-task set (88.95%, p = 0.748). A non-significant result of this kind is not a formal test of equivalence. The same subset significantly beats a 25-feature RFE baseline drawn from the full feature space (p = 0.014).

The three selection methods disagreed on individual features (Jaccard 0.06 to 0.36 at k = 10) but agreed on which tasks matter. The Hill fit places half-saturation at 4 features and 90% saturation at 36 for Random Forest. Tasks 1 through 7, the simple graphic and copying tasks, are consistently the least useful; a double-check subset built from the lowest-importance tasks reached only 83.43%, 5.52 points below full ($p = 0.0001$). Leave-one-out cross-validation at k = 10 (83.3%) matched the 30-seed split estimate (83.4%) to within rounding. This concerns the k = 10 per-feature model and is separate from the 10-task TOP10_MEAN analysis.

Evaluated within the DARWIN cohort, a 10-task candidate set performed comparably to the full 25-task set on accuracy and AUC while using 60% fewer tasks. These results are internal and cohort-specific. The main remaining question is external, cross-cohort validation of the 10-task set, which is the subject of a planned prospective follow-up study.

Declarations

Competing interests:

The authors declare no competing interests. Funding: This research received no external funding. Corresponding author: Sanjay Singhvi, CompuTerra Inc. (email: sanjay@computerra.com). Use of AI tools: The authors used an AI assistant (Anthropic Claude) to support manuscript editing, and revision; the authors are responsible for all content, including the study design, experiments, analyses, and conclusions, and reviewed and verified the manuscript in full. Ethics and scope: This study is a secondary analysis of the publicly available DARWIN handwriting dataset and involved no new recruitment, intervention, or human-participant data collection. The findings represent exploratory internal evaluation and should not be interpreted as clinical diagnostic validation. Author contributions: R.S. and S.S. designed the study, implemented the experiments, analyzed the results, and wrote the manuscript.

Data and Code Availability

Data availability:

The DARWIN dataset is publicly available from the UCI Machine Learning Repository at the DOI cited in the References.

Code availability: Analysis code and configuration files will be made available upon preprint publication.

References

1. Alzheimer’s Association (2022) 2022 Alzheimer’s disease facts and figures. Alzheimer’s Dement 18(4):700–789. https://doi.org/10.1002/alz.12638

2. Arevalo-Rodriguez I, Smailagic N, Roque-Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, Pedraza OL, Cosp B, X., Cullum S (2015) Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane Database of Systematic Reviews, 2015(3), CD010783. https://doi.org/10.1002/14651858.CD010783.pub2
3. Chen C, Thebaud T, Butala A, Dehak N, Moro-Velazquez L, Oh ES (2024) Cognitive assessment through writing tasks. Alzheimer's Dement 20:e093186. https://doi.org/10.1002/alz.093186
4. Cilia ND, De Stefano C, Fontanella F, di Scotto A (2018) An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. Procedia Comput Sci 141:466–471. https://doi.org/10.1016/j.procs.2018.10.141
5. Cilia ND, D'Alessandro T, De Stefano C, Fontanella F, Molinara M (2021) From online handwriting to synthetic images for Alzheimer's disease detection using a deep transfer learning approach. IEEE J Biomedical Health Inf 25(12):4243–4254. https://doi.org/10.1109/JBHI.2021.3101982
6. Cilia ND, De Gregorio G, De Stefano C, Fontanella F, Marcelli A, Parziale A (2022) Diagnosing Alzheimer's disease from online handwriting: A novel dataset and performance benchmarking. Eng Appl Artif Intell 111:104822. https://doi.org/10.1016/j.engappai.2022.104822
7. De Stefano C, Fontanella F, Impedovo D, Pirlo G, di Scotto A (2019) Handwriting analysis to support neurodegenerative diseases diagnosis: A review. Pattern Recognit Lett 121:37–45. https://doi.org/10.1016/j.patrec.2018.05.013
8. Fernandes CP, Montalvo G, Caligiuri M, Pertsinakis M, Guimarães J (2023) Handwriting changes in Alzheimer's disease: A systematic review. J Alzheimer's Disease 96(1):1–11. https://doi.org/10.3233/JAD-230438
9. Fontanella F (2022) DARWIN [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55D0K
10. Ho NTN, Gonzalez P, Gogovi GK (2025) Writing the signs: An explainable machine learning approach for Alzheimer's disease classification from handwriting. Healthc Technol Lett 12(1):e70006. https://doi.org/10.1049/htl2.70006
11. Impedovo D, Pirlo G (2019) Dynamic handwriting analysis for the assessment of neurodegenerative diseases: A pattern recognition perspective. IEEE Rev Biomed Eng 12:209–220. https://doi.org/10.1109/RBME.2018.2840679
12. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
13. Tahami Monfared AA, Byrnes MJ, White LA, Zhang Q (2022) Alzheimer's disease: Epidemiology and clinical progression. Neurol Therapy 11(2):553–569. https://doi.org/10.1007/s40120-022-00338-8

Figures

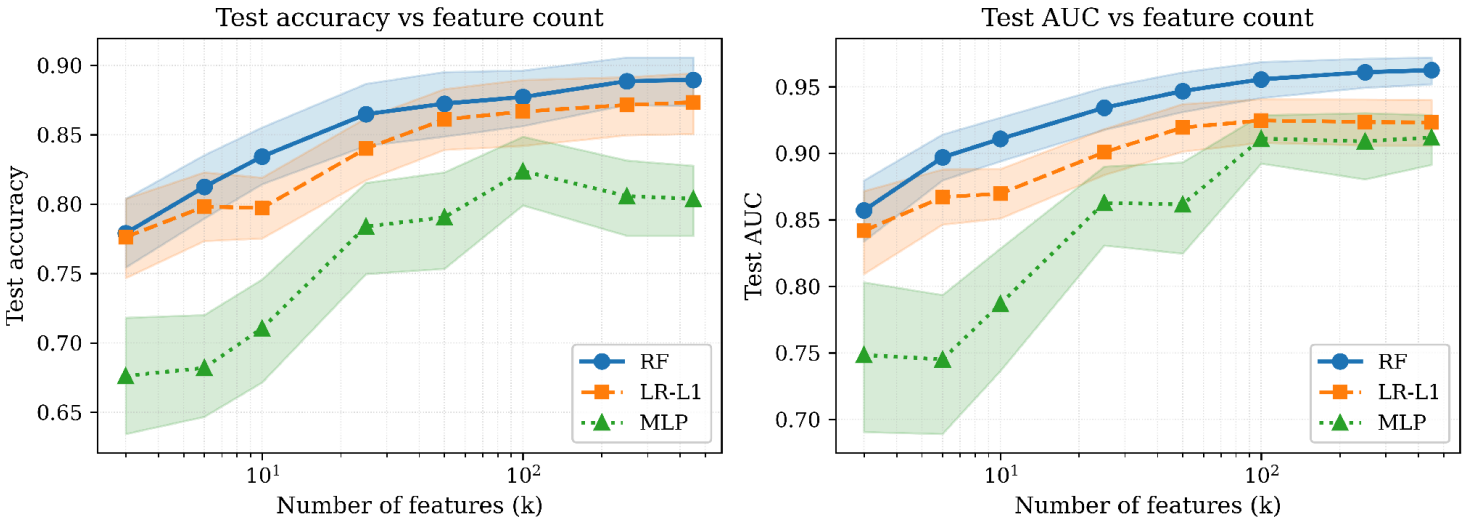


Figure 1

Test Accuracy and AUC as a Function of the Number of Selected Features

Note. Test accuracy (left) and AUC (right) versus the number of features, k, for Random Forest (circles, solid blue), L1-regularized Logistic Regression (squares, dashed black), and a small Multilayer Perceptron (triangles, dotted gray). Lines are the mean over 30 stratified 80/20 splits; shaded bands are bootstrap 95% confidence intervals. The x-axis is on a log scale. Different markers, line styles, and colors keep the three models distinguishable in grayscale. Random Forest leads throughout, and the accuracy curve is strongly concave: at k = 25 the confidence intervals overlap the k = 450 intervals for all three models, but at k = 10 they do not.

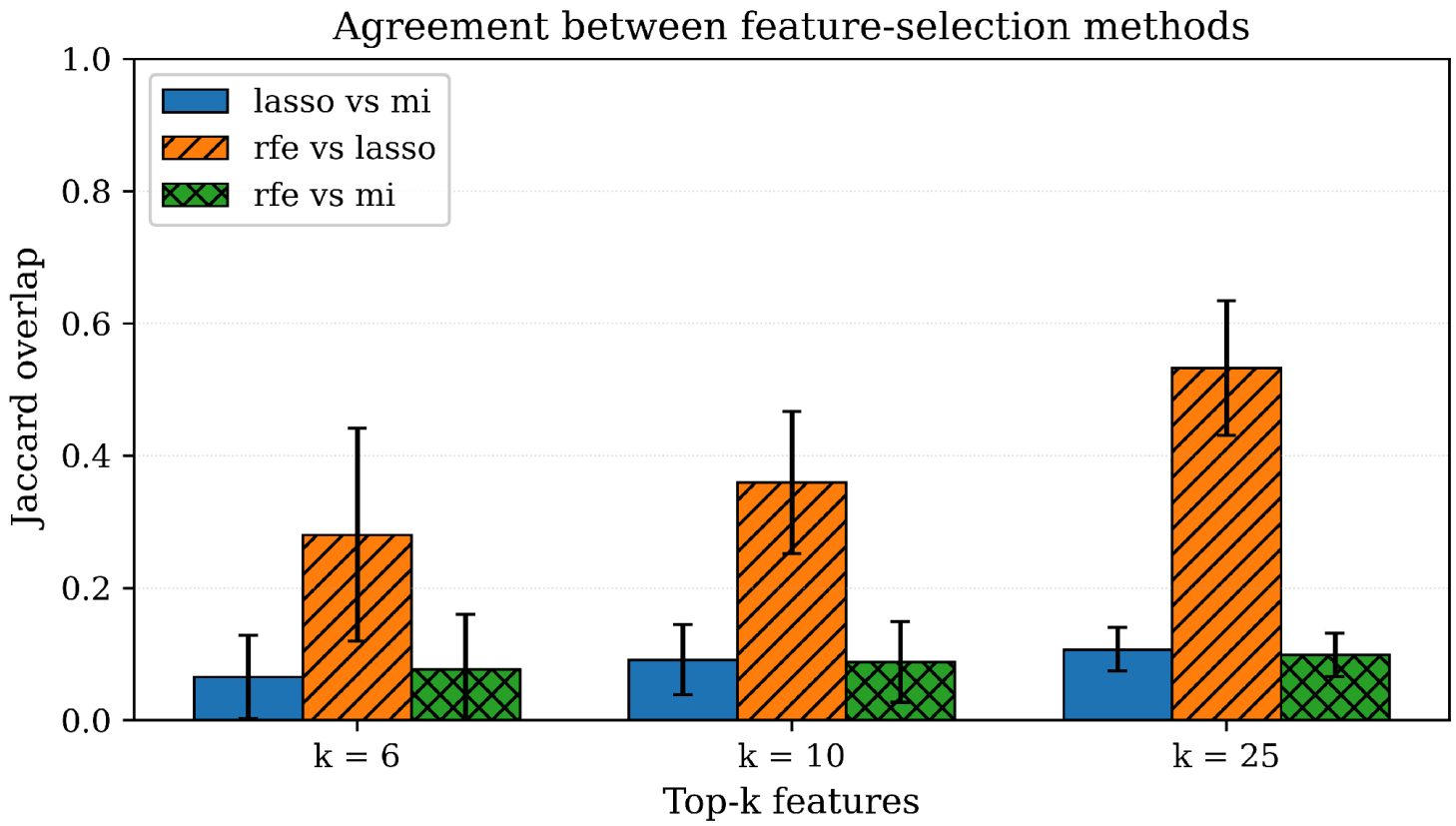


Figure 2

Agreement Between Feature-Selection Methods

Note. Jaccard overlap between the top-k feature sets chosen by RFE, LASSO, and mutual information (MI) at k = 6, 10, and 25, averaged over 30 seeds with standard-deviation error bars. RFE and LASSO, both Logistic-Regression-based, agree moderately. MI disagrees with both at the feature level.

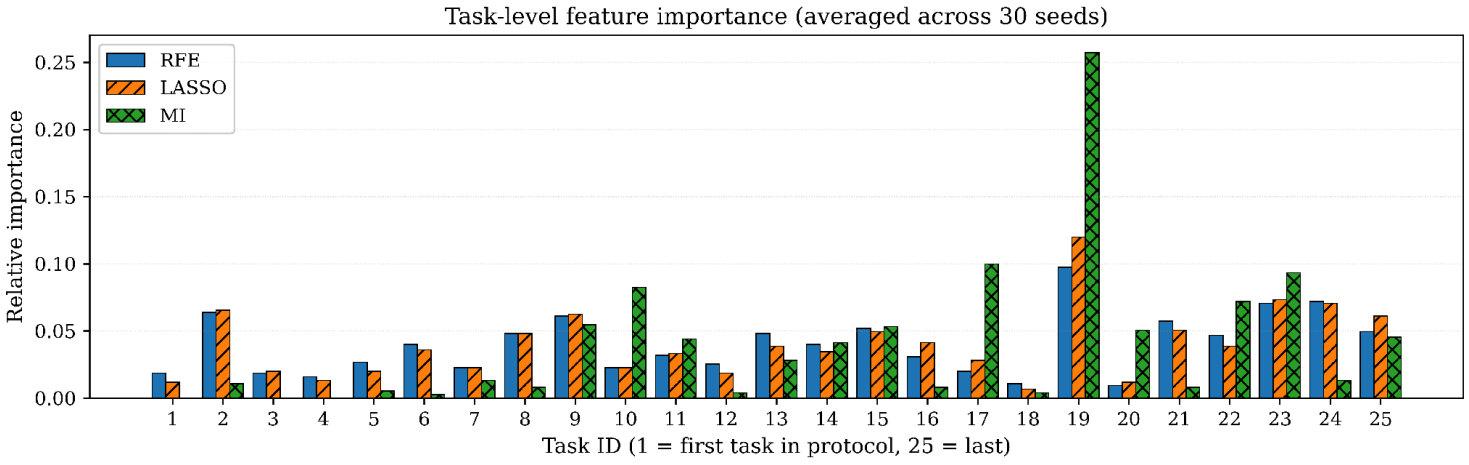


Figure 3

Task-Level Importance Under Three Feature-Selection Methods

Note. Task-level importance for each of the 25 DARWIN tasks under the three feature-selection methods. For each method, the 18 features per task were summed and normalized within that method, then averaged over 30 seeds at k = 25. Despite low feature-level agreement (Figure 2), all three methods rank task 19 as most important and tasks 23, 9, 25, 15, 24, and 22 as jointly informative, while tasks 1 through 7 are consistently low.

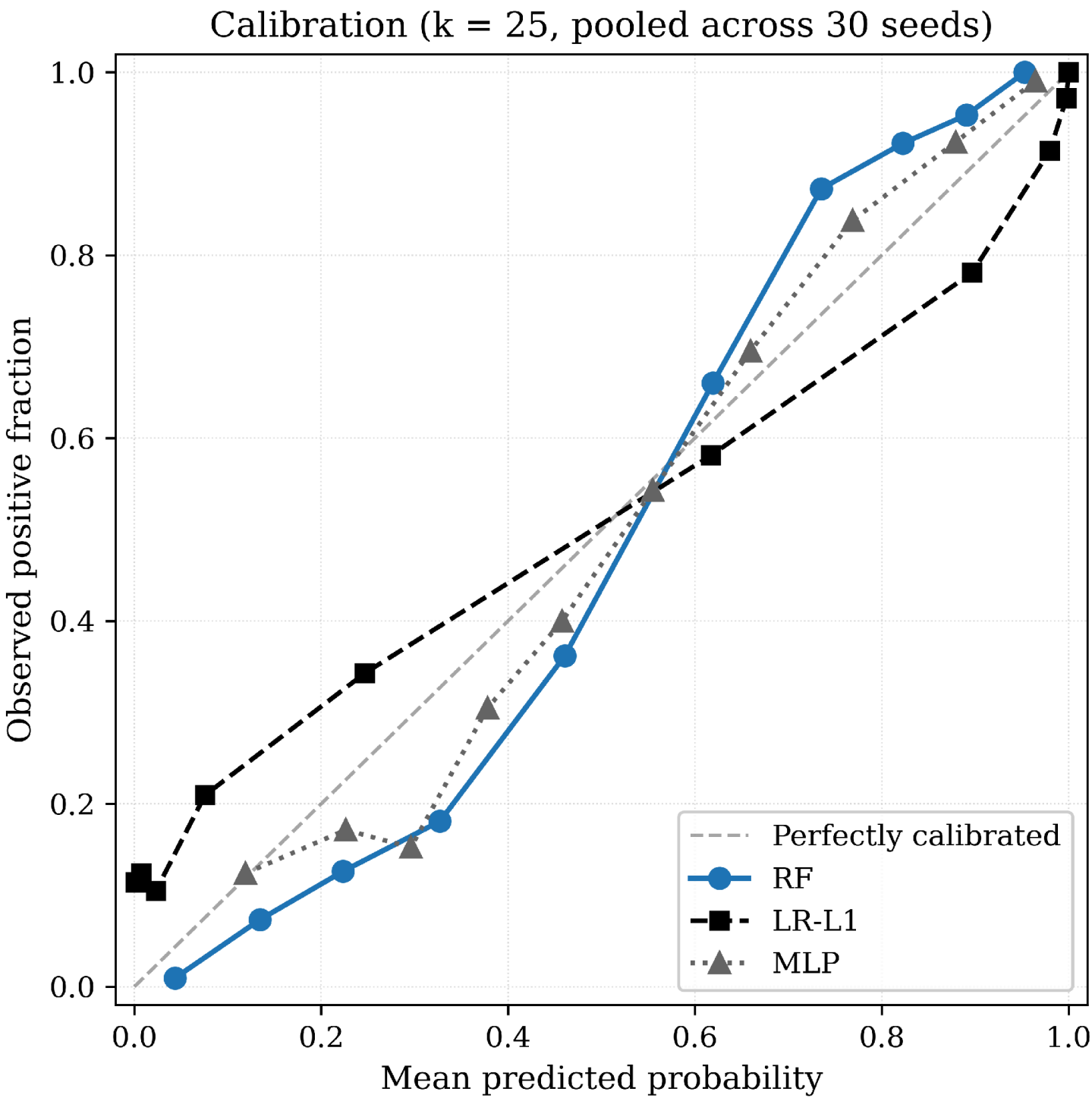


Figure 4

Reliability Diagram for the Three Classifiers at k = 25

Note. Reliability diagram for the three classifiers at k = 25. Predicted probabilities were sorted into 10 quantile-based bins, and the observed fraction of positives is plotted against the mean predicted probability in each bin. Pooled across 30 seeds (1050 predictions per model). The diagonal marks perfect calibration.

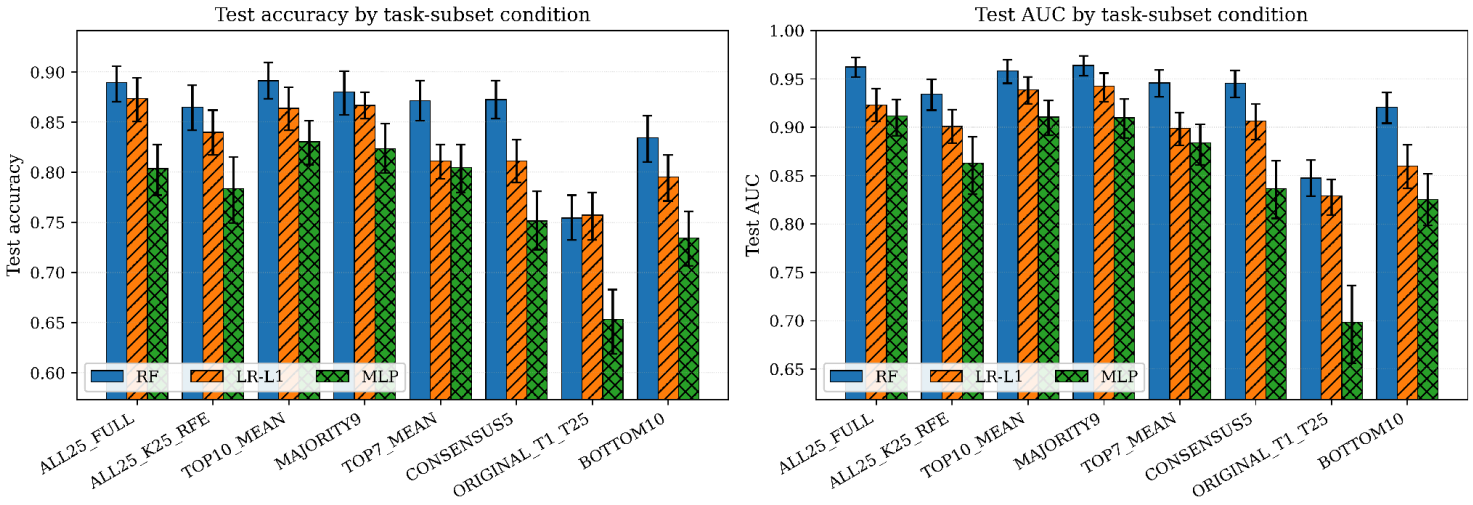


Figure 5

Test Accuracy and AUC for Each Task Subset

Note. Random Forest, L1-Logistic Regression, and MLP test accuracy (left) and AUC (right) for each of the eight task subsets, shown as the mean over 30 stratified 80/20 splits with bootstrap 95% confidence intervals as error bars. ALL25_FULL is the reference. TOP10_MEAN, MAJORITY9, TOP7_MEAN, and CONSENSUS5 all reach RF accuracy within sampling noise of the reference. ORIGINAL_T1_T25 (the first and last tasks) and BOTTOM10 (the 10 least-informative tasks) are clearly worse, confirming the task ranking. Different markers, fill colors, and hatch patterns keep the three models readable in grayscale.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx