Task Complexity Matters in Patient-Level Offline Handwriting Classification for Alzheimer Screening

1st Stefano Antonio Amico
Department of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
stefano.amico@infoamico.com

2nd Ludovica Beritelli
Dep. of Mathematics and Computer Science
University of Catania
Catania, Italy
ludovica.beritelli@phd.unict.it

3rd David Panebianco
Department of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
david.panebianco@phd.unict.it

4th Roberta Avanzato
Dep. of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
roberta.avanzato@unict.it

5th Francesco Beritelli
Dep. of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
francesco.beritelli@unict.it

***Abstract*—This paper presents a patient-level study on first-level screening for Alzheimer's disease from offline handwriting images. The analysis is based on DARWIN-I handwriting samples and compares two Ultralytics YOLO-based image classifiers, YOLO11 and YOLO26, under a strict patient-level split designed to prevent identity confounding. The dataset is partitioned into 140 training subjects, 17 validation subjects, and 17 test subjects, corresponding to 1744, 208, and 221 images, respectively. Two experimental scenarios are evaluated. In the baseline setting, based on short handwriting tasks, YOLO26 improves global accuracy from 55% to 61% and patient-class F1-score from 52% to 57% with respect to YOLO11, but recall remains limited at 50%. In the long-writing setting, restricted to Task 14 and Task 25, both models reach 59% accuracy; however, their clinical behavior diverges sharply. YOLO11 attains 75% precision but only 33% patient recall, whereas YOLO26 reaches 72% recall and 65% F1-score. These findings show that task complexity is not a secondary variable: sustained sentence-copying tasks appear to expose disease-related graphomotor alterations more clearly than short traces. The study should be interpreted as a proof of concept, because the cohort is small, image-level reporting is used, and no external validation is available.**

***Index Terms*—Alzheimer's Disease, Handwriting Analysis, Digital Biomarkers, Computer Vision, Patient-Level Split, YOLO11, YOLO26, DARWIN-I.**

I. Introduction

Dementia is a major public-health challenge, and Alzheimer's disease is its most common cause [1], [2]. Current clinical and research frameworks emphasize that AD assessment combines clinical syndrome definition with biomarker evidence when available [2], [3]. Early screening is clinically relevant because it can support timely referral, follow-up, and treatment planning. However, gold-standard diagnostic pathways may require specialist assessment, imaging, or fluid biomarkers, which are not always suitable as scalable first-level tools. This has motivated the study of non-invasive digital biomarkers.

Handwriting is a promising candidate because it reflects the interaction of motor control, visuospatial organization, linguistic planning, and executive function [4], [5]. In neurodegenerative conditions, subtle impairments in these functions may affect stroke morphology, spatial regularity, continuity, and task execution strategy. Most prior studies have exploited online handwriting acquired through tablets or smart pens, where velocity, pressure, and in-air time are explicitly available [6], [7]. The present work instead studies a more scalable offline setting: the input is the final static image of the handwriting trace.

The study investigates whether modern image classifiers can distinguish healthy controls (HC) from patients (PT) using DARWIN-I handwriting images. The work is methodologically centered on a strict patient-level split, which avoids the leakage effect that would occur if different images from the same subject appeared in both training and testing [8]. The study compares two Ultralytics classification architectures, YOLO11 and YOLO26, and evaluates them in two clinically meaningful scenarios: short, low-complexity tasks and long sentence-copying tasks.

This paper makes three contributions:

- it presents a patient-level computer-vision protocol for Alzheimer's disease screening from offline handwriting images;
- it reports the comparative behavior of YOLO11 and YOLO26 on short and long handwriting tasks;
- it shows that task complexity substantially changes clinical utility, especially patient recall.

II. Related Work

Handwriting has long been studied as a window on cognitive and motor impairment. Slavin *et al.* reported altered consistency of handwriting movements in dementia of the Alzheimer type, while Schröter *et al.* found reduced fine-motor

regularity in probable Alzheimer's disease and MCI [9], [10]. Werner *et al.* showed that handwriting process variables can help discriminate mild Alzheimer's disease and mild cognitive impairment from healthy aging [4]. Later work confirmed that kinematic and pressure features may differ across MCI, AD, and healthy controls [11]. More recently, reviews have consolidated both the opportunities and methodological challenges of handwriting-based biomarkers for neurodegenerative diseases [5], [7].

A recurring methodological theme in this literature is that the diagnostic value of handwriting depends strongly on the task being performed. Short copying tasks may isolate local stroke morphology, whereas longer text-copying tasks require sustained attention, working memory, visuospatial alignment, and fine motor control over a longer interval. Recent task-selection and word-level analyses on Alzheimer handwriting data further support the idea that the eliciting task changes the discriminative information available to the model [12], [13]. This is why the experiment preserves the distinction between short low-complexity tasks and long sentence-copying tasks instead of aggregating all images into a single undifferentiated test condition.

The DARWIN dataset introduced a public benchmark for Alzheimer's disease classification from online handwriting, and its UCI record makes it accessible for broader experimentation [6], [14]. Cilia *et al.* also demonstrated that online handwriting trajectories can be transformed into synthetic images suitable for deep transfer learning, and subsequent offline-image work further supports the relevance of image-based handwriting analysis [15], [16].

From the machine-learning perspective, deep convolutional networks are attractive because they learn visual features directly from pixels rather than from hand-engineered descriptors [17]. Although the YOLO family was originally introduced for object detection [18], the Ultralytics framework now supports image classification as well [19]–[21]. In this work, each handwriting image is treated as a single two-class classification sample.

III. Materials and Methods

A. Dataset and Task Groups

The study is based on DARWIN-I, an image representation derived from the DARWIN handwriting protocol [6], [14]. The original cohort contains 174 subjects: 85 HC and 89 PT. Unlike online experiments, the present setting discards explicit temporal and pressure channels and retains only the final two-dimensional handwritten trace.

Two task groups are considered. The first group defines the *baseline* setting and contains short tasks (single letters, numbers, symbols, or short words): Task 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 18, 22, and 23. The second group isolates the two long sentence-copying tasks (Task 14 and Task 25), which were selected to test whether prolonged visuomotor effort and denser text structure reveal stronger graphomotor alterations.

TABLE I
Patient-level split used in the experiment.

Subset	Patients	HC/PT	Images
Training	140	69/71	1744
Validation	17	8/9	208
Test	17	8/9	221

B. Patient-Level Split

A central methodological choice is the patient-level split. All images belonging to the same participant are assigned to a single subset, so the network is never evaluated on unseen images from a writer already observed during training. This prevents identity confounding and makes the reported performance clinically more credible [8]. The experiment implements an 80–10–10 split at subject level, with stratification to preserve the proportion of HC and PT subjects in each subset. The resulting distribution is shown in Table I. The important point is the order of operations: patient identifiers are extracted first, and only afterwards are images assigned to training, validation, or test sets. This prevents a frequent but serious error in multi-sample medical datasets, namely evaluating on images from the same individual already seen during training.

C. Models and Evaluation Protocol

The two compared architectures are YOLO11 and YOLO26, both used in image-classification mode through the Ultralytics framework [19]–[21]. The models were trained for 100 epochs in a Colab-based environment. The goal was not only to maximize overall accuracy, but also to understand which architecture provides more clinically useful behavior in a screening context. For this reason, the evaluation emphasizes patient-class recall and false-negative rate in addition to global accuracy.

Let PT be the positive class. Performance is summarized by accuracy, precision, recall, and F1-score. Recall is especially important because false negatives correspond to missed pathological cases.

IV. Results

A. Baseline Results on Short Tasks

The baseline experiment evaluates the models on short handwriting tasks. As reported in Table II, YOLO26 outperforms YOLO11 on all metrics, but the margins are moderate. In particular, patient recall increases from 45% to 50%, which is still insufficient for a reliable first-level screening tool.

The baseline test confusion matrix for YOLO11 is reported in Fig. 3. The model correctly identifies 69 HC images and 53 PT images, but it misclassifies 64 pathological samples as healthy controls. This confirms the clinically problematic false-negative tendency. For the remaining experimental conditions, the evaluation is summarized through precision, recall,

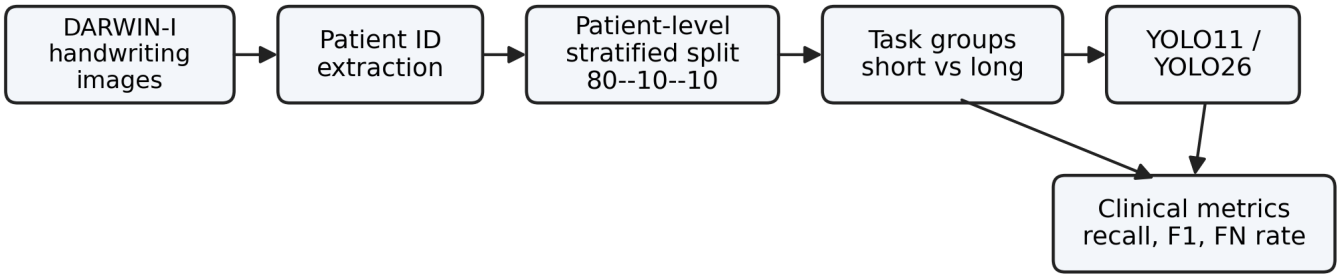

Fig. 1. Condensed overview of the experimental pipeline: image-based DARWIN-I samples are grouped by patient, split at subject level, divided into short- and long-task subsets, and then evaluated with YOLO11 and YOLO26.

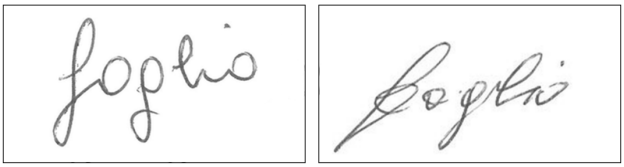

Fig. 2. Representative short-task handwriting samples from the DARWIN-I image setting (word "foglio").

TABLE II
BASELINE PERFORMANCE ON SHORT TASKS. PRECISION, RECALL, AND F1-SCORE REFER TO THE PATIENT CLASS.

Model	Accuracy	Precision	Recall	F1
YOLO11	55%	60%	45%	52%
YOLO26	61%	67%	50%	57%

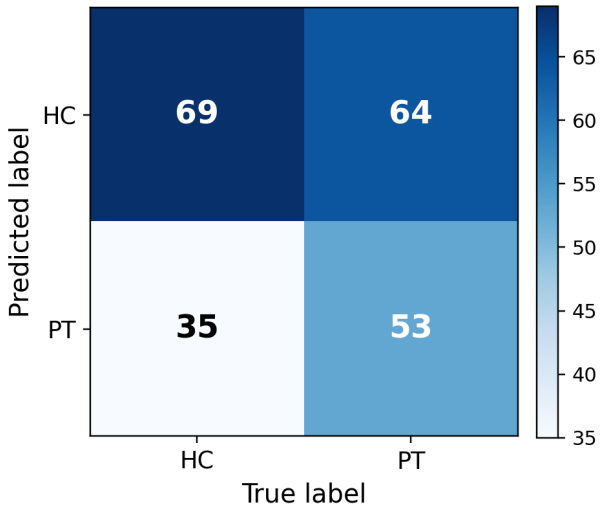

Fig. 3. Baseline test confusion matrix for YOLO11.

F1-score, and false-negative rate, as these aggregate measures capture the clinically relevant trade-offs more directly than per-cell counts for a dataset of this size.

B. Task-Complexity Effect: Long Sentence-Copying Tasks

The central result emerges when the analysis is restricted to the long sentence-copying tasks (Task 14 and Task 25). In this scenario, the two architectures have the same global accuracy (59%), but their clinical profiles diverge radically. As shown in Table III, YOLO11 is conservative: it reaches 75% precision but only 33% recall. In contrast, YOLO26 becomes markedly more sensitive, reaching 72% recall and 65% F1-score.

The grouped comparison in Fig. 4 makes the main message of the study explicit. On short tasks, YOLO26 yields a consistent but limited improvement. On long tasks, the gain becomes clinically meaningful because recall rises by 39 percentage points with respect to YOLO11. This suggests that long tasks provide richer visual evidence of graphomotor degradation, while the deeper architecture is better able to exploit that evidence.

For a screening application, the clinically relevant comparison is not only whether the classifier is correct on average,

TABLE III
PERFORMANCE ON LONG SENTENCE-COPYING TASKS (TASK 14 AND TASK 25). PRECISION, RECALL, AND F1-SCORE REFER TO THE PATIENT CLASS.

Model	Accuracy	Precision	Recall	F1
YOLO11	59%	75%	33%	46%
YOLO26	59%	59%	72%	65%

but how often it misses patients. Fig. 5 therefore re-expresses the reported recalls as false-negative rates. In the long-task condition, the false-negative rate decreases from 67% for YOLO11 to 28% for YOLO26. False-negative rate is therefore used as the primary screening-oriented summary, since missed pathological cases have a direct clinical cost in terms of delayed referral.

V. DISCUSSION

The results support three observations. First, the patient-level split is essential. In handwriting analysis, a naive image-

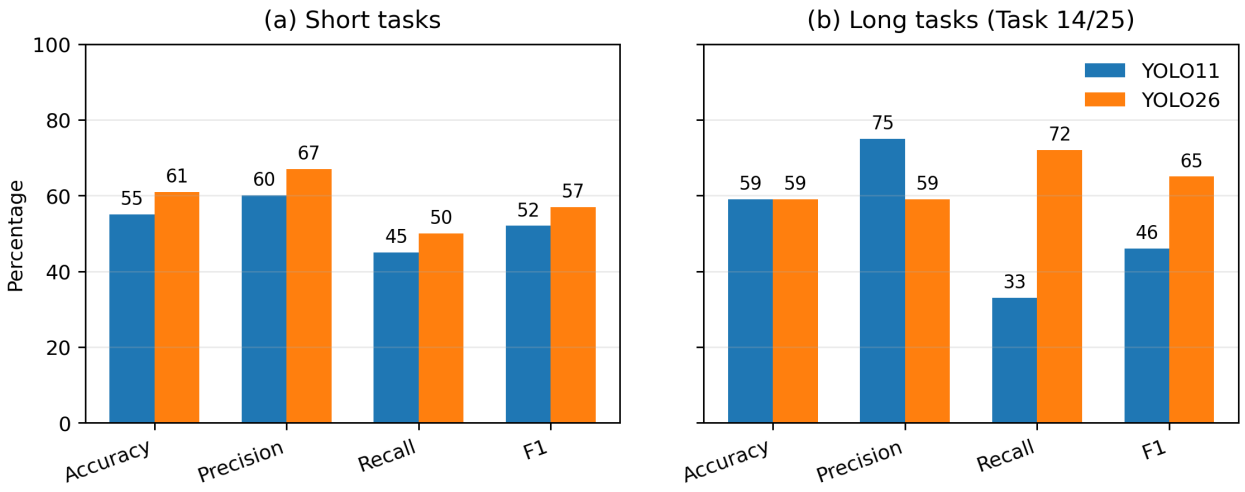

Fig. 4. Metric comparison between YOLO11 and YOLO26 across the two experimental scenarios. The most relevant shift appears in the long-task setting (right), where YOLO26 sharply increases patient recall while maintaining the same overall accuracy reported for YOLO11.

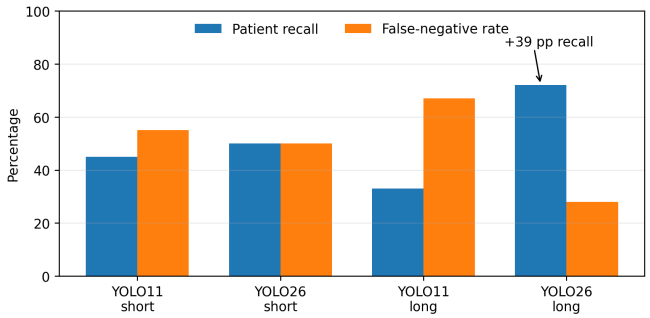

Fig. 5. Screening-oriented summary of patient recall and false-negative rate. Long sentence-copying tasks substantially improve the sensitivity profile of YOLO26.

level random split could produce overly optimistic metrics because writing style is highly individual. Second, overall accuracy is inadequate for judging clinical usefulness. In the long-task scenario, both architectures reach 59% accuracy, yet their utility differs substantially because recall behaves in opposite ways. Third, task design must be considered part of the machine-learning system. The classifier is not simply analyzing handwriting "in general"; it is analyzing handwriting under a particular cognitive and motor load.

A plausible explanation for the long-task effect is that sentence-copying requires sustained visuomotor control, denser spatial organization, and longer execution time. These factors may amplify disease-related alterations such as instability, hesitations, irregular spacing, or stroke degradation. The clinical chart in Fig. 5 is consistent with this interpretation: when the task becomes denser and more demanding, the deeper architecture shifts toward substantially higher sensitivity.

From a clinical point of view, the trade-off between precision and recall must be interpreted cautiously. A screening-oriented tool is expected to prioritize sensitivity, because false negatives delay clinical referral. In this respect, YOLO26 on long tasks is preferable to YOLO11, even though its precision is lower. Nevertheless, a 72% recall still leaves a substantial number of missed cases, which means that the current system should be regarded as a support tool rather than as a standalone diagnostic instrument.

The observed precision–recall shift also suggests that model selection cannot be separated from task selection. If the input protocol consists only of short isolated traces, the deeper architecture has limited opportunity to exploit disease-related visual evidence. If the protocol includes sentence-copying, the input contains more layout decisions, stroke transitions, and fatigue-sensitive segments. The classifier, the task battery, and the clinical operating point therefore form a coupled system. For this reason, future comparisons should report results not only by architecture, but also by task family and by aggregation rule across tasks.

VI. Limitations and Future Work

The study remains preliminary. The patient-level test split contains only 17 subjects, and the reported metrics are image-level rather than subject-level. Consequently, confidence intervals would likely be wide, and clinical deployment would require aggregation across multiple tasks for each participant. Furthermore, the experiment does not include external validation or calibration against an independent clinical cohort.

Future work should therefore proceed in four directions. First, repeated patient-level cross-validation and subject-level aggregation should be reported, so that sensitivity and specificity can be estimated at participant level rather than at image level. Second, explainability methods such as Grad-CAM should be used to verify that the networks attend to plausible stroke regions rather than to spurious background cues [22]. Third, inspired by broader handwriting-screening research, the offline image pipeline could be integrated with online channels such as pressure, velocity, and in-air time,

TABLE IV
RECOMMENDED ROADMAP FOR STRENGTHENING THE STUDY.

Step	Purpose
Repeated patient-level CV	Estimate variability and confidence intervals.
Subject-level voting	Aggregate multiple task predictions per participant.
Grad-CAM/occlusion maps	Verify attention on clinically plausible stroke regions.
External cohort	Test robustness beyond the source dataset.
Online/offline fusion	Add velocity, pressure, and in-air time when available.

thereby enriching the biomarker space [4], [6], [23]. Finally, the image-based setting could be extended with synthetic-image or sim-to-real strategies, following the direction already explored for handwriting trajectories in the literature [15], [24]. These future directions are not part of the present experimental evidence and should be validated with separate protocols.

Table IV summarizes the most important technical additions needed before a stronger clinical claim can be made. The priority is to move from image-level results to subject-level endpoints, because clinical screening decisions are made for individuals, not for isolated task images. A second priority is model interpretation: without visual explanations or occlusion tests, it is impossible to exclude the possibility that the network relies on acquisition artifacts or writer-specific background features rather than on handwriting morphology.

Another important extension would be external validation on data acquired with different pens, scanners, or clinical protocols. Offline handwriting models can be highly sensitive to acquisition artifacts such as line thickness, page contrast, and compression noise. A robust screening pipeline should therefore be stress-tested across heterogeneous acquisition conditions before any clinical adoption is considered. The image-only setting remains attractive because it is inexpensive, but its scalability should not be confused with immediate diagnostic reliability. The most appropriate near-term use is therefore a referral-support tool: high-risk cases can be prioritized for specialist assessment, while negative image-only results should not override clinical symptoms or caregiver reports.

A. Reproducibility Considerations

The experimental evidence reported in this study is limited to offline image classification on DARWIN-I, patient-level partitioning, and the comparison between YOLO11 and YOLO26 under the two task groups. Dynamic sensing variables such as pressure, velocity, and in-air time are not used in the reported experiments; they are considered only as future extensions. This distinction is important because the offline and online settings rely on different input modalities and should be validated with separate protocols before any combined claim is made.

In this study, the partition was defined at subject level before any image assignment, following the reporting principles for clinical prediction models outlined in TRIPOD+AI [25]. Reporting the total number of images is insufficient when each participant contributes multiple samples. The most useful release would include the patient identifiers assigned to training, validation, and test subsets, or a deterministic script that reconstructs the same partition from the file names. This would allow independent researchers to verify that no participant appears across multiple subsets and would make architecture comparisons independent of random split variation.

A further reproducibility issue concerns the decision threshold. The current experiment reports class labels and standard metrics, but a practical screening system should expose class probabilities and allow the operating point to be selected according to the clinical objective. A first-level screening tool may prefer a sensitivity-oriented threshold, accepting more false positives to reduce missed patients. A confirmatory support tool would require a higher-specificity threshold. Future work should therefore report receiver operating characteristic curves, precision–recall curves, and threshold-dependent sensitivity–specificity trade-offs.

B. Threats to Validity

Several threats to validity remain even after the patient-level split. First, the long-task subset is necessarily smaller than the full baseline setting, so the apparently large recall gain should be interpreted with caution. A few additional subjects could materially change the metric values. Second, the reported labels refer to images, not individuals. A patient who produces multiple task images may have mixed predictions across tasks; this uncertainty should be handled by subject-level probability aggregation.

Third, the offline representation removes temporal information. This is both the strength and the limitation of the proposed approach. It enables low-cost deployment from scanned or photographed tests, but it cannot directly measure speed, pressure, pen lifts, or hesitation time. These variables are clinically meaningful and have been widely discussed in online handwriting research. The most realistic long-term system may therefore be hybrid: offline images for broad screening, followed by online acquisition for borderline or high-risk cases.

VII. CONCLUSION

This paper presented a patient-level offline handwriting classification study for Alzheimer's disease screening. The comparison between YOLO11 and YOLO26 shows that architecture depth matters, but task complexity matters even more. On short tasks, YOLO26 improves the baseline only moderately. On long sentence-copying tasks, it reaches 72% patient recall and clearly outperforms YOLO11 in clinically relevant sensitivity. The findings support the use of task-aware computer-vision pipelines as low-cost screening aids, while also highlighting the need for larger cohorts, subject-level reporting, external validation, stronger reporting transparency, and explainability before translation to real clinical practice.

These findings highlight that evaluation rigor in handwriting-based AI must extend equally to model architecture, patient-level partitioning, and clinical task design. None of these factors can be treated as secondary without risk of misleading performance estimates.

REFERENCES

[1] World Health Organization, "Dementia: Fact sheet," Online, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] C. R. J. Jack *et al.*, "Nia-aa research framework: Toward a biological definition of alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[3] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. J. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[4] P. Werner, S. Rosenblum, G. Bar-On, J. Heinik, and A. D. Korczyn, "Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment," *The Journals of Gerontology: Series B*, vol. 61, no. 4, pp. P228–P236, 2006.

[5] C. De Stefano, F. Fontanella, D. Impedovo, G. Pirlo, and A. Scotto di Freca, "Handwriting analysis to support neurodegenerative diseases diagnosis: A review," *Pattern Recognition Letters*, vol. 121, pp. 37–45, 2019.

[6] N. D. Cilia, G. De Gregorio, C. De Stefano, F. Fontanella, A. Marcelli, and A. Parziale, "Diagnosing alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking," *Engineering Applications of Artificial Intelligence*, vol. 111, p. 104822, 2022.

[7] M. Moetesum, M. Diaz, U. Masroor, I. Siddiqi, and G. Vessio, "A survey of visual and procedural handwriting analysis for neuropsychological assessment," *Neural Computing and Applications*, vol. 34, pp. 9561–9578, 2022.

[8] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 15:1–15:21, 2012.

[9] M. J. Slavin, J. G. Phillips, J. L. Bradshaw, K. A. Hall, and I. Presnell, "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, no. 1, pp. 20–25, 1999.

[10] A. Schröter, R. Mergl, K. Bürger, H. Hampel, H.-J. Möller, and U. Hegerl, "Kinematic analysis of handwriting movements in patients with alzheimer's disease, mild cognitive impairment, depression and healthy subjects," *Dementia and Geriatric Cognitive Disorders*, vol. 15, no. 3, pp. 132–142, 2003.

[11] J. Garre-Olmo, M. Faúndez-Zanuy, K. López-de Ipiña, L. Calvó-Perxas, and O. Turró-Garriga, "Kinematic and pressure features of handwriting and drawing: Preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls," *Current Alzheimer Research*, vol. 14, no. 9, pp. 960–968, 2017.

[12] V. Gattulli, N. D. Cilia, F. Fontanella, and C. De Stefano, "Handwriting task-selection based on the analysis of the performance of machine learning models for alzheimer's disease diagnosis," in *CEUR Workshop Proceedings*, vol. 3521, 2023.

[13] N. D. Cilia, C. De Stefano, F. Fontanella, and S. M. Siniscalchi, "How word semantics and phonology affect handwriting of alzheimer's patients: A machine learning based analysis," arXiv:2307.04762, 2023. [Online]. Available: https://arxiv.org/abs/2307.04762

[14] UCI Machine Learning Repository, "DARWIN: Diagnosis AlzheimeR WIth haNdwriting," Online dataset record, 2022. [Online]. Available: https://archive.ics.uci.edu/dataset/732/darwin

[15] N. D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella, and M. Molinara, "From online handwriting to synthetic images for alzheimer's disease detection using a deep transfer learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4243–4254, 2021.

[16] N. D. Cilia, T. D'Alessandro, C. De Stefano, and F. Fontanella, "Offline handwriting image analysis to predict alzheimer's disease via deep learning," in *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 2807–2813.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[19] Ultralytics, "Image classification with ultralytics yolo," Documentation, 2026. [Online]. Available: https://docs.ultralytics.com/tasks/classify/

[20] ——, "Ultralytics yolo11," Documentation, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/

[21] ——, "Ultralytics yolo26," Documentation, 2025. [Online]. Available: https://docs.ultralytics.com/models/yolo26/

[22] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[23] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Analysis of in-air movement in handwriting: A novel marker for parkinson's disease," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 405–411, 2014.

[24] I. Bazarbekov *et al.*, "Sim-to-real domain adaptation for early alzheimer's detection from handwriting kinematics using hybrid deep learning," *Sensors*, vol. 26, no. 1, p. 298, 2026.

[25] G. S. Collins *et al.*, "Tripod+ai statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024.