

Article

A Hybrid Meta-Classifer Framework for Alzheimer's Disease Classification Using Handwriting Analysis

Nadhir Djeflal ^{1,2} , Salem Titouni ² , Abdallah Hedir ^{1,3,*} , Mohamed Salah Bouaouina ¹, Mounir Amir ¹  and Idris Messaoudene ² 

- ¹ Faculty of Electrical Engineering and Computer Science, University of Mouloud Mammeri, Tizi Ouzou 15000, Algeria; nadhir.djeflal@ummto.dz (N.D.); mohamed-salah.bouaouina@ummto.dz (M.S.B.); mounir.amir@ummto.dz (M.A.)
- ² Laboratoire d'Electronique et Télécommunications Avancées, University Mohamed El Bachir El Ibrahimi of Bordj Bou Arreridj, Bordj Bou Arreridj 34000, Algeria; salem.titouni@univ-bba.dz (S.T.); idris.messaoudene@univ-bba.dz (I.M.)
- ³ Advanced High Voltage Engineering Centre, School of Engineering, Cardiff University, Queen's Buildings, The Parade, Cardiff CF24 3AA, UK
- * Correspondence: abdallah.hedir@ummto.dz

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder for which accurate and accessible screening approaches remain an important research objective. This study proposes a hybrid meta-classifier framework for handwriting-based AD classification using the DARWIN (Diagnosis Alzheimer WITH haNdwriting) dataset, which comprises 174 participants, including 89 individuals with AD and 85 healthy controls, and contains 450 handwriting-related features derived from 25 tasks. The proposed framework combines a one-dimensional convolutional neural network (1D-CNN) for automated feature learning with a heterogeneous ensemble of XGBoost, support vector machine (SVM), and random forest classifiers. The predictions of the base classifiers are subsequently integrated by an AdaBoost-based meta-classifier to generate the final classification decision. The framework achieved a pooled cross-validation accuracy of 98.28%, with a mean fold-level accuracy of $97.73 \pm 1.27\%$, a precision of $96.90 \pm 2.84\%$, a recall of $98.89 \pm 2.49\%$, an F1-score of $97.84 \pm 1.21\%$, and an AUC of 0.944 ± 0.023 across the five outer folds. In addition, evaluation on an independent handwriting-signature cohort achieved an accuracy of 91.76%, with a sensitivity of 90.74% and a specificity of 93.55%. These results indicate that the proposed framework has strong discriminative capability across the evaluated datasets. Nevertheless, further validation on larger and more diverse independent cohorts is required before conclusions regarding clinical applicability can be drawn. Future work will investigate multimodal integration and the application of explainable artificial intelligence techniques to improve the interpretability of the proposed framework.



Academic Editor: Jing Jin

Received: 6 August 2026

Revised: 10 September 2026

Accepted: 12 September 2026

Published: 18 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Alzheimer's disease; handwriting analysis; deep learning; CNN; AdaBoost; hybrid meta-classifier; XGBoost; support vector machine; random forest; digital biomarkers

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the most prevalent cause of dementia among the elderly population worldwide. It is clinically characterized by gradual deterioration of cognitive functions, including episodic memory loss, executive dysfunction, language impairment, and impaired visuospatial and fine

motor skills [1,2]. In recent years, non-invasive and automated approaches have shown great potential for the early detection of neurodegenerative disorders [3–6]. However, despite significant advances in neuroimaging techniques and biomarker discovery, the early diagnosis of AD remains challenging. Conventional diagnostic approaches often rely on expensive and invasive clinical assessments, such as magnetic resonance imaging (MRI) and positron emission tomography (PET). These approaches are frequently constrained by high costs, limited accessibility, and invasiveness, thus preventing large-scale screening and early intervention efforts [7,8].

In response to these limitations, there has been growing interest in the development of non-invasive, cost-effective, and scalable digital biomarkers for the early detection of AD. Among the emerging modalities, handwriting analysis has attracted significant interest as a promising cognitive-motor indicator of neurodegenerative alterations [9–11]. Handwriting is a complex psychomotor activity that involves the integration of multiple cognitive and motor processes, including working memory, language processing, attention, visuospatial coordination, fine motor control, and sensorimotor feedback. Consequently, even subtle alterations in handwriting kinematics, dynamics, and pressure patterns may reflect underlying neural degradation associated with AD, particularly in the early stages when clinical symptoms are still mild [12,13].

Recent empirical studies have consistently demonstrated a strong correlation between specific handwriting features and the clinical manifestation of AD. These features include total time, in-air and on-paper time, mean speed and acceleration, pressure variability, jerk, tremor indices, and spatial dispersion, among others [14,15]. By capturing these micro-level motor signatures, handwriting-based diagnostic frameworks offer the potential to detect disease-related changes before they become apparent in traditional cognitive screening tests.

Within the biomedical signal processing and machine learning communities, a wide spectrum of computational methodologies have been explored to enhance the diagnostic reliability and classification accuracy of handwriting-based AD detection systems. These approaches can be broadly categorized into three main paradigms: (1) handcrafted feature extraction combined with traditional machine learning classifiers (e.g., support vector machines, random forests, k-nearest neighbors); (2) deep neural network architectures, particularly convolutional neural networks (CNNs), which automatically learn hierarchical representations from raw or minimally preprocessed data; and (3) ensemble or hybrid learning frameworks that combine multiple classifiers to leverage their complementary strengths [16–22].

Despite substantial progress reported in the literature, several critical challenges persist. First, many current models demonstrate limited robustness when applied to heterogeneous data conditions, such as variations in recording devices, task instructions, or demographics of participants [23]. Second, cross-domain generalizability remains a concern, as models trained on specific datasets often fail to generalize to independent clinical populations [24]. Third, the development of adaptive decision-fusion mechanisms that can effectively integrate predictions from diverse base classifiers is still an open area of research [25]. Collectively, these limitations underscore the need for further investigation to improve the reliability, scalability, and clinical applicability of handwriting-based AD diagnostic systems.

In this context, the present study introduces a novel hybrid methodology for the early detection of Alzheimer's disease using handwriting analysis, evaluated on the publicly available DARWIN dataset [14]. The DARWIN dataset contains handwriting samples from 174 participants, including 89 individuals diagnosed with AD and 85 healthy controls. Participants completed 25 handwriting tasks spanning memory recall, dictation, graphic

drawing, and copying exercises. A total of 18 handwriting features are extracted per task, yielding a comprehensive feature space of 450 dimensions.

Although handwriting analysis may provide useful information for the development of accessible screening tools, the present study should be interpreted as a handwriting-based classification study rather than as a prospective demonstration of preclinical or early diagnosis. The DARWIN dataset contains participants already labelled as having AD or as healthy controls. Therefore, the objective of the proposed framework is to investigate the ability of handwriting-related features to discriminate between these groups under the adopted experimental protocol. Further longitudinal and clinically representative studies would be required to establish whether such approaches can support the detection of AD at earlier or preclinical stages.

The proposed methodology integrates three principal components: (1) a deep convolutional neural network (CNN) designed to extract high-dimensional latent features from the handwriting data; (2) a hybrid ensemble model combining XGBoost, support vector machine (SVM), and random forest classifiers, selected for their complementary capabilities in modeling complex nonlinear relationships, handling high-dimensional feature spaces, and reducing prediction variance; and (3) an AdaBoost meta-classifier that synthesizes the predictions from the base models through an adaptive boosting procedure, iteratively focusing on misclassified instances to refine overall decision boundaries. This multi-layered architecture aims to address key challenges related to class imbalance, feature complexity, and model robustness, designed to address key challenges related to feature complexity, model diversity, and diagnostic robustness.

The main methodological contributions of this study are summarized as follows:

- We develop a unified framework for Alzheimer's disease assessment from structured handwriting descriptors, integrating multiple handwriting tasks into a fixed-dimensional feature representation rather than relying on raw handwriting images.
- We introduce a one-dimensional CNN-based representation learning module that transforms the 450-dimensional handwriting feature vector into a compact 64-dimensional latent representation, enabling the downstream classifiers to operate on a learned representation of the structured handwriting information.
- We propose an ensemble meta-learning strategy that combines XGBoost, SVM, and Random Forest through an AdaBoost meta-classifier. The meta-features are generated using an inner-fold prediction procedure to avoid using in-sample predictions from the downstream base classifiers during meta-classifier training.
- We employ a nested cross-validation framework in which preprocessing and model-development operations are restricted to the corresponding outer training data, while the outer test fold remains unseen until final evaluation. This provides a structured separation between model development and performance assessment.
- We further assess the generalizability of the DARWIN-trained framework on an independent Signature cohort acquired under a substantially different handwriting protocol. The external cohort is used exclusively for independent evaluation and is not involved in model training, fine-tuning, feature selection, hyperparameter optimization, or threshold determination.

The remainder of this paper is organized as follows. Section 2 reviews related work on handwriting-based AD diagnosis. Section 3 describes the DARWIN dataset in detail. Section 4 presents the proposed methodology, including preprocessing, model architectures, and meta-classifier training. Section 5 reports and discusses the experimental results, including performance comparisons with existing methods. Finally, Section 6 concludes the paper with a summary of findings and directions for future research.

2. Related Work

Recent advances in digital health, biomedical computing, and artificial intelligence have significantly increased research interest in handwriting-based diagnostic systems for AD. Handwriting is increasingly recognized as a sensitive psychomotor biomarker because it simultaneously reflects cognitive planning, motor coordination, visuospatial processing, executive functioning, and sensorimotor integration. Consequently, subtle impairments caused by neurodegenerative disorders can manifest through measurable alterations in handwriting dynamics, including velocity, acceleration, pressure variability, tremor patterns, and spatial organization [9,10].

Existing literature has explored a broad range of machine learning (ML), deep learning (DL), and ensemble learning techniques for exploiting these characteristics to support early AD diagnosis. Broadly, prior studies can be grouped into three major categories: (1) feature extraction and signal-processing approaches, (2) deep learning-based handwriting analysis methods, and (3) hybrid and ensemble learning frameworks [14–21].

2.1. Feature Extraction and Signal Processing Approaches

Early investigations of handwriting-based AD diagnosis primarily relied on handcrafted feature extraction and classical signal-processing methodologies. These studies focused on identifying psychomotor abnormalities reflected in temporal, spatial, and kinematic handwriting characteristics, including writing duration, pressure variability, velocity, acceleration, jerk, tremor indices, and spatial dispersion [12,13].

More recent experimental studies have demonstrated the clinical relevance of online handwriting signals for AD detection. Ref. [14] introduced the DARWIN dataset and established one of the first comprehensive benchmarking frameworks for handwriting-based Alzheimer's diagnosis using machine learning techniques. Their work demonstrated that online handwriting dynamics contain highly discriminative information capable of differentiating AD patients from healthy controls. Similarly, ref. [26] highlighted the effectiveness of convolutional architectures in automatically learning discriminative latent representations from handwriting signals while reducing the dependency on manually engineered descriptors.

Several studies subsequently incorporated traditional machine learning classifiers into handwriting analysis pipelines. Ref. [15] proposed a multi-classifier framework combining different ML algorithms for Alzheimer's diagnosis using online handwriting signals, demonstrating improved classification robustness compared with single-model approaches. Ref. [18] investigated one-class classifiers to address imbalanced clinical datasets and reported strong diagnostic performance using anomaly-detection strategies. Other researchers explored feature optimization mechanisms to improve classification performance. Ref. [19] introduced a hybrid machine learning model integrated with Particle Swarm Optimization (PSO) for selecting discriminative handwriting features related to Alzheimer's disease.

Despite these advances, important methodological limitations remain insufficiently addressed. First, most feature-engineering approaches rely heavily on handcrafted descriptors whose effectiveness depends strongly on expert-driven feature selection and preprocessing strategies. Such dependency may limit reproducibility and reduce adaptability across heterogeneous handwriting acquisition conditions. Second, many studies are evaluated exclusively on relatively small datasets, particularly the DARWIN dataset, increasing the risk of overfitting and reducing statistical robustness. Third, several classical machine learning pipelines exhibit limited cross-domain generalization because handwriting characteristics are highly sensitive to demographic variability, acquisition devices, writing protocols, and motor conditions unrelated to AD. Furthermore, handcrafted feature extraction may

fail to capture complex nonlinear interactions embedded within handwriting dynamics, particularly subtle latent patterns associated with early-stage cognitive decline. These limitations motivate the development of more adaptive and data-driven architectures capable of combining automated feature learning with robust ensemble decision mechanisms.

2.2. Deep Learning-Based Handwriting Analysis

The emergence of deep learning has significantly transformed handwriting-based AD diagnosis by enabling automated hierarchical feature extraction from raw or minimally processed handwriting data. Unlike conventional machine learning pipelines that depend heavily on handcrafted descriptors, deep neural networks can directly learn complex latent representations from high-dimensional handwriting signals. CNNs, in particular, have demonstrated a strong capability in identifying subtle spatial and temporal patterns associated with neurodegenerative disorders. Ref. [26] proposed a CNN-based framework for early Alzheimer's detection using handwriting signals and demonstrated that deep convolutional architectures outperform several conventional machine learning techniques in capture latent handwriting dynamics. Similarly, ref. [16] introduced a machine learning-powered handwriting analysis system for early Alzheimer's diagnosis and highlighted the effectiveness of deep latent representations in identifying subtle cognitive impairments that may not be observable only through handcrafted features. More recently, ref. [17] proposed a hybrid cognitive handwriting framework integrating deep learning with advanced pre-processing strategies to improve classification robustness and diagnostic precision under heterogeneous handwriting conditions.

These studies collectively demonstrate that deep learning models can effectively learn discriminative representations without requiring extensive manual feature engineering. In addition, deep architectures offer several advantages, including automated feature learning, scalability, nonlinear modeling capability, and improved adaptability to complex handwriting patterns. CNN-based approaches are particularly effective in capturing local spatial dependencies and hidden correlations among kinematic handwriting variables, which are often difficult to model using traditional statistical techniques.

Nevertheless, despite their promising performance, current deep-learning approaches to handwriting-based AD classification face important methodological limitations. Many studies are trained and evaluated on relatively small datasets, including the DARWIN dataset, which comprises only 174 participants. The combination of a limited sample size and a high-dimensional feature space may increase the risk of overfitting and lead to optimistic performance estimates. Consequently, reported classification results should be interpreted in the context of the adopted validation protocol, and evaluation of independent cohorts remains essential to assess generalizability.

Furthermore, deep learning models generally require substantial computational resources for training and inference, potentially limiting their deployment in low-resource or real-time clinical settings. Existing studies also rarely investigate robustness against domain shifts caused by differences in acquisition devices, handwriting protocols, demographic variability, or comorbid motor impairments. Consequently, although CNN-based handwriting analysis has shown considerable promise for early AD detection, further research is needed to improve the generalizability, interpretability, computational efficiency, and clinical reproducibility of the model before large-scale real-world deployment can be achieved.

2.3. Hybrid and Ensemble Learning Frameworks

To overcome the limitations of standalone classifiers, recent research has increasingly explored hybrid and ensemble learning frameworks for handwriting-based AD diagnosis.

Ensemble methodologies aim to combine multiple classifiers in order to exploit their complementary strengths, thereby improving robustness, reducing variance, and enhancing classification stability under complex and heterogeneous handwriting conditions. These approaches are particularly attractive in biomedical applications where single classifiers may fail to capture the multidimensional relationships embedded within cognitive-motor handwriting patterns. Ref. [21] proposed a triple ensemble learning architecture for Alzheimer's detection using handwriting analysis and demonstrated that integrating heterogeneous classifiers significantly improves predictive performance compared with individual machine learning models. Similarly, ref. [20] designed and validated a hybrid machine learning framework combining multiple learning paradigms for AD diagnosis, reporting enhanced robustness and improved discrimination capability across diverse handwriting samples. Boosting-based strategies have also shown promising diagnostic performance. Ref. [24] investigated boosting classification algorithms for Alzheimer's diagnosis and demonstrated that adaptive iterative learning mechanisms can improve classification accuracy under challenging classification scenarios. Furthermore, ref. [23] explored handwriting task-selection strategies and revealed that certain handwriting tasks contribute disproportionately to diagnostic performance, highlighting the importance of task-sensitive learning mechanisms in handwriting-based AD detection systems.

More recent studies have attempted to integrate deep learning with ensemble decision strategies in order to combine automated feature extraction with robust classification mechanisms. Such hybrid architectures seek to leverage the representation-learning capability of CNNs alongside the stability and interpretability of classical machine learning classifiers such as SVM, Random Forest, and XGBoost. These systems generally demonstrate improved robustness against noisy or high-dimensional handwriting data and often outperform standalone deep learning or traditional machine learning approaches. In addition to improving classification performance, recent studies have also explored explainable machine learning for handwriting-based Alzheimer's disease classification. For example, Ho et al. applied SHapley Additive exPlanations (SHAP) to identify the handwriting features contributing to model predictions, highlighting the importance of temporal and pressure-related characteristics across different handwriting tasks [27].

Despite these promising developments, several important methodological limitations remain evident in current ensemble-based handwriting diagnosis systems. First, many existing hybrid frameworks employ static decision-fusion mechanisms in which classifier contributions remain fixed during inference. Such static fusion strategies may not dynamically adapt to heterogeneous handwriting conditions, difficult clinical cases, or variations in severity cognitive impairment. Second, the increased architectural complexity of ensemble systems often leads to a higher computational overhead, reducing their practicality for real-time or embedded clinical deployment. Third, several studies report exceptionally high classification accuracies without performing external validation or cross-dataset evaluation, raising concerns about model generalizability and reproducibility.

Another important limitation concerns interpretability. Although ensemble models frequently improve predictive performance, the integration of multiple classifiers and boosting mechanisms can further reduce transparency in the diagnostic decision process. This lack of explainability can hinder clinical trust and regulatory acceptance in healthcare applications. In addition, most existing hybrid frameworks remain strongly dependent on the DARWIN dataset and have not been validated across diverse demographic populations, acquisition devices, or multilingual handwriting conditions. Consequently, there remains a critical need for adaptive ensemble architectures capable of combining deep feature extraction, dynamic decision fusion, computational efficiency, and clinically interpretable reasoning mechanisms to support robust real-world Alzheimer's disease screening.

Table 1 summarizes representative handwriting-based AD diagnosis studies, covering feature selection methods, deeplearning architectures, and ensemble strategies. The comparison highlights not only reported performance, but also key methodological limitations, including data set dependency, reliance on oversampling, high computational complexity, and limited robustness of clinical handwriting data.

Table 1. Comparison of Handwriting-Based Alzheimer’s Disease Diagnosis Studies.

Ref.	Model(s)	Dataset	Main Contribution	Performance	Limitation
[14]	Benchmark ML classifiers	DARWIN	Introduced the DARWIN handwriting dataset and established baseline benchmarking for AD diagnosis using online handwriting features	94.28% Accuracy	Limited generalization across heterogeneous populations and devices
[15]	Multi-classifier ML framework	Online handwriting dataset	Combined multiple classifiers to improve robustness in handwriting-based AD detection	91% Accuracy	Relied heavily on handcrafted feature engineering
[18]	One-Class Classifiers	DARWIN	Investigated anomaly-based learning for Alzheimer’s diagnosis under limited pathological data conditions	97.12% Accuracy	Sensitive to dataset distribution and class imbalance
[19]	Hybrid ML + PSO	Handwriting dataset	Optimized feature selection using Particle Swarm Optimization to improve classification performance	90% Accuracy	High computational complexity due to optimization procedures
[20]	Hybrid Machine Learning Model	Handwriting dataset	Developed and validated a hybrid framework integrating multiple ML paradigms for AD detection	High diagnostic robustness	Limited explainability and dependence on handcrafted features
[21]	Triple Ensemble Learning Model	DARWIN	Proposed a triple ensemble framework to improve predictive stability and classification accuracy	97.14% Accuracy	Increased computational overhead and static decision fusion
[22]	CNN-based framework	DARWIN	Demonstrated the effectiveness of CNNs for automated latent handwriting feature extraction	90.4% Accuracy	Risk of overfitting due to limited dataset size

Table 1. Cont.

Ref.	Model(s)	Dataset	Main Contribution	Performance	Limitation
[23]	Task-selection classification framework	DARWIN	Identified the most discriminative handwriting tasks for Alzheimer's diagnosis	Improved task efficiency	Limited adaptability to unseen handwriting conditions
[24]	Boosting classification algorithms	DARWIN	Applied boosting strategies to improve classification accuracy and robustness	85% Accuracy	Lower performance compared with ensemble and deep learning approaches
[16]	ML-powered handwriting analysis	Handwriting dataset	Applied AI-driven handwriting analytics for early Alzheimer's screening	Competitive performance	Dataset dependency and limited cross-domain validation
[17]	Hybrid Deep Learning Framework	Handwriting dataset	Combined cognitive handwriting analysis with hybrid AI techniques for improved AD diagnosis	Improved robustness and precision	High model complexity and computational cost
Proposed Work	CNN + XGBoost + SVM + RF + AdaBoost Meta-Classifier	DARWIN	Integrated deep feature extraction with heterogeneous ensemble learning and adaptive boosting for robust AD detection	98.28% Accuracy, 0.94 AUC	Requires further cross-domain clinical validation and real-world deployment testing

2.4. Recent Transformer-Based Approaches

Recent advances in deep learning have increasingly explored Transformer-based architectures for neurodegenerative disease assessment. Bone et al. [28] proposed a Vision Transformer (ViT) framework for automated dementia screening using clock-drawing test images, demonstrating the potential of attention-based models for capturing complex cognitive patterns. Similarly, Gong et al. [29] introduced a Hybrid Transformer model that integrates handwriting-based 2D images and 1D dynamic handwriting signals, achieving state-of-the-art performance on the DARWIN dataset through multimodal attention mechanisms. More recently, Sharma and Acharya [30] developed a lightweight Vision Transformer with transfer learning for Alzheimer's disease severity assessment, highlighting the effectiveness of Transformer-based feature representation and interpretability techniques. These studies indicate a growing interest in Transformer architectures for cognitive impairment assessment; however, their higher computational complexity and data requirements remain important challenges. In contrast, the proposed framework combines CNN-based feature extraction with heterogeneous ensemble learning and adaptive meta-classification, providing a computationally efficient and highly accurate alternative for handwriting-based Alzheimer's disease detection.

Recent advances in medical artificial intelligence have increasingly focused on improving model interpretability, multimodal information integration, and robust deep learning architectures. For example, Li et al. proposed an explainable graph network that integrates

image-derived information with medical knowledge to improve both diagnostic performance and interpretability in knee osteoarthritis assessment [31]. Multimodal information fusion has also been investigated for medical image analysis, with recent work proposing decision-level fusion of multimodal MRI information for robust cartilage segmentation and clinically confirmed knee osteoarthritis assessment [32]. In addition, recent deep learning approaches have explored advanced architectures for medical image segmentation. SwM-rNet, for instance, integrates an improved Swin Transformer with a multi-scale residual module to enhance global feature representation and local detail extraction for multi-target knee tissue segmentation [33]. These studies highlight the growing importance of explainability, complementary information fusion, and advanced representation learning in medical AI. In contrast, the present study focuses on structured handwriting-derived features and investigates a CNN-based representation learning and ensemble classification framework for Alzheimer's disease assessment, with independent external validation on a separate signature dataset.

2.5. Research Gap and Motivation

Although existing studies consistently demonstrate the effectiveness of handwriting analysis for early Alzheimer's disease detection, several important research gaps remain unresolved. First, many current systems remain highly dataset-dependent and exhibit poor generalization across diverse demographic populations, writing conditions, and acquisition devices [22,23]. Second, numerous approaches rely heavily on handcrafted features and oversampling techniques, potentially limiting scalability and robustness. Third, while deep learning methods achieve strong predictive performance, they often suffer from overfitting and limited interpretability when trained on relatively small clinical datasets.

Moreover, current ensemble frameworks generally employ static fusion strategies that lack adaptive learning capabilities. Consequently, they may fail to effectively handle heterogeneous handwriting patterns and difficult instances during classification. Addressing these limitations requires more robust hybrid architectures capable of integrating automated deep feature extraction with adaptive ensemble learning strategies.

Motivated by these limitations, this study proposes a hybrid meta-classifier framework that differs from existing handwriting-based Alzheimer's disease detection methods in three key aspects. First, a CNN is employed to automatically learn latent representations from the high-dimensional handwriting feature space, reducing reliance on handcrafted descriptors. Second, a heterogeneous ensemble composed of XGBoost, SVM, and Random Forest exploits complementary learning mechanisms to improve classification stability and robustness. Third, an AdaBoost-based meta-classifier performs adaptive decision fusion, enabling dynamic refinement of classifier outputs rather than relying on conventional static ensemble strategies. This combination of deep feature learning, heterogeneous ensemble modeling, and adaptive boosting has not been previously explored in a unified framework for Alzheimer's disease detection using the DARWIN handwriting dataset.

3. Dataset Description

3.1. First Dataset: DARWIN Dataset

The DARWIN dataset has been meticulously assembled to aid in the early identification of AD through the analysis of handwriting characteristics. It contains contributions from 174 individuals, including 89 diagnosed with Alzheimer's disease and 85 who are healthy. The dataset is derived from 25 distinct tasks divided into three groups: Memory and dictation, Graphic tasks, and copy tasks, as outlined in Table 2. Each task measures 18 different handwriting features, detailed in Table 3, culminating in a total of 450 distinct

features related to handwriting analysis. Information on the average demographics of the participants used to compile the dataset is detailed in Table 4.

Table 2. List of tasks performed.

Task #	Description	Category
1	Signature drawing	M
2	Join two points with a horizontal line, continuously for four times	G
3	Join two points with a vertical line, continuously for four times	G
4	Retrace a circle (6 cm of diameter) continuously for four times	G
5	Retrace a circle (3 cm of diameter) continuously for four times	G
6	Copy the letters 'l', 'm' and 'p'	C
7	Copy the letters on the adjacent rows	C
8	Write cursively a sequence of four lowercase letter 'l', in a single smooth movement	C
9	Write cursively a sequence of four lowercase cursive bigram 'le', in a single smooth movement	C
10	Copy the word "foglio"	C
11	Copy the word "foglio" above a line	C
12	Copy the word "mamma"	C
13	Copy the word "mamma" above a line	C
14	Memorize the words "telefono", "cane", and "negozio" and rewrite them	M
15	Copy in reverse the word "bottiglia"	C
16	Copy in reverse the word "casa"	C
17	Copy six words (regular, non-regular, non words) in the appropriate boxes	C
18	Write the name of the object shown in a picture (a chair)	M
19	Copy the fields of a postal order	C
20	Write a simple sentence under dictation	M
21	Retrace a complex form	G
22	Copy a telephone number	C
23	Write a telephone number under dictation	M
24	Draw a clock, with all hours and put hands at 11:05 (Clock Drawing Test)	G
25	Copy a paragraph	C

Table 3. List of features associated with each task.

Feature #	Feature Name	Feature Description
1	Total Time (TT)	Total time spent to perform the entire task
2	Air Time (AT)	Time spent to perform in-air movements
3	Paper Time (PT)	Time spent to perform on-paper movements
4	Mean Speed on-paper (MSP)	Average speed of on-paper movements. Speed is the variation of displacement with respect to time.
5	Mean Speed in-air (MSA)	Average speed of in-air movements.
6	Mean Acceleration on-paper (MAP)	Average acceleration of on-paper movements. Acceleration is the variation of speed with respect to time.
7	Mean Acceleration in-air (MAA)	Average acceleration of in-air movements.
8	Mean Jerk on-paper (MJP)	Average jerk of on-paper movements. Jerk is the variation of acceleration with respect to time.
9	Mean Jerk in-air (MJA)	Average jerk of in-air movements.
10	Pressure Mean (PM)	Average of the pressure levels exerted by the pen tip.
11	Pressure Var (PV)	Variance of the pressure levels exerted by the pen tip.

Table 3. *Cont.*

Feature #	Feature Name	Feature Description
12	GMRT on-paper (GMRTp)	Generalization of the Mean Relative Tremor (MRT) as defined in Pereira et al. (2015) [34] and computed for on-paper movements.
13	GMRT in-air (GMRTa)	Generalization of the Mean Relative Tremor computed on in-air movements.
14	Mean GMRT (GMRT)	Average of GMRTp and GMRTa.
15	Pendowns Number (PWN)	Counts the total number of pendowns recorded during the execution of the entire task
16	Max X Extension (XE)	Maximum extension recorded along the X axis.
17	Max Y Extension (YE)	Maximum extension recorded along the Y axis.
18	Dispersion Index (DI)	The Dispersion Index measures how the handwritten trace is ‘dispersed’ on the entire piece of paper; in other words, it measures how much of the sheet is covered.

Table 4. Average participant demographic information. Standard deviations are indicated in brackets.

Participant Class	Age	Women	Men
Patients	71.5 (9.5)	46	44
Healthy people	68.9 (12)	51	39

3.2. Second Dataset: Handwritten Signature Dataset

To further assess the generalizability of the proposed hybrid meta-classifier beyond the DARWIN cohort, an independent publicly available handwriting dataset was considered for external validation. The dataset was introduced by Preti et al. [35] in the study entitled “Does cognitive decline influence signing?” and is publicly available through the Open Science Framework (OSF) repository. Unlike the DARWIN dataset, which consists of quantitative handwriting descriptors extracted from multiple graphomotor tasks, this dataset focuses specifically on handwritten signatures and their evolution over time in individuals with neurodegenerative diseases and neurologically healthy controls.

3.2.1. Dataset Composition

The dataset comprises 134 participants, including 54 individuals diagnosed with Alzheimer’s disease (AD), 49 individuals diagnosed with frontotemporal dementia (FTD), and 31 neurologically healthy controls (HC). The diagnoses of AD and FTD were established according to multiparametric clinical criteria. In particular, AD diagnoses were based on the International Working Group 2 (IWG-2) criteria, whereas FTD diagnoses followed the revised diagnostic criteria reported in the corresponding clinical literature [35].

For the purpose of the present study, only the AD and HC groups were retained in order to maintain the same binary classification objective as that adopted with the DARWIN dataset. Consequently, the FTD participants were excluded from the external validation analysis. The resulting external validation cohort therefore consisted of 85 participants, including 54 AD patients and 31 healthy controls.

Table 5 summarizes the composition of the original cohort and the subset retained for the present study.

Table 5. Composition of the publicly available handwriting-signature dataset.

Group	Original Cohort	Used in This Study	Percentage
Alzheimer’s disease (AD)	54	54	63.53%
Frontotemporal dementia (FTD)	49	0	–
Healthy controls (HC)	31	31	36.47%
Total	134	85	100%

The resulting AD-to-HC ratio is relatively balanced, with 54 AD participants and 31 healthy controls. Although the Signature dataset is smaller than the DARWIN dataset, it provides an independent population and therefore offers an opportunity to evaluate whether the proposed model retains its discriminative capability under a different data collection protocol.

3.2.2. Participant Characteristics

The clinical cohort was retrospectively recruited from the Neurodegenerative Diseases Unit of the Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico in Milan, Italy. All participants underwent neurological and neuropsychological evaluation. The AD group consisted of 54 participants with a mean age of 73.8 ± 6.4 years, while the healthy control group consisted of 31 participants with a mean age of 73.1 ± 10.7 years. The mean educational level was 9.3 ± 4.3 years for the AD group and 10.6 ± 4.7 years for the HC group. The mean Mini-Mental State Examination (MMSE) score was 22.2 ± 4.8 for the AD group and 28.76 ± 1.9 for the HC group.

Table 6 provides the demographic and clinical characteristics of the two groups retained for the external validation.

Table 6. Demographic and clinical characteristics of the AD and HC groups in the external validation cohort. Values are reported as mean $\pm$ standard deviation unless otherwise indicated.

Characteristic	AD	HC
Number of participants	54	31
Age (years)	73.8 ± 6.4	73.1 ± 10.7
Male/Female	15/39	14/17
Education (years)	9.3 ± 4.3	10.6 ± 4.7
MMSE score	22.2 ± 4.8	28.76 ± 1.9

The age distributions of the two groups are relatively similar, whereas the AD group presents a lower mean MMSE score, consistent with the presence of cognitive impairment. The study also excluded participants with a clinical history of visual, auditory, language, or motor disorders, as well as muscular rigidity or hypo/hypertrophy, in order to reduce potential confounding effects on handwriting performance.

3.2.3. Signature Acquisition and Longitudinal Design

Table 7 presents a distinctive characteristic of this dataset is its retrospective longitudinal design. Two handwritten signatures were available for each participant. For individuals with AD or FTD, the first signature was obtained from an identity card and corresponded to a period preceding the diagnosis, whereas the second signature was obtained from an informed-consent document for cerebrospinal fluid examination and corresponded to a period following diagnosis. Only pairs of signatures separated by at least one year were considered. The mean time interval between the two signatures was approximately 5.7 ± 2.3 years, with a range from 1 to 18 years.

For healthy controls, one signature was obtained from an identity card and the second signature was collected during the study on an A4 sheet. The mean interval between the two signatures was 4.03 ± 2.98 years, ranging from 1 to 13 years.

Table 7. Signature acquisition protocol in the external validation dataset.

Characteristic	Clinical Groups	Healthy Controls
Number of signatures per participant	2	2
First signature	Identity card	Identity card
Second signature	CSF informed-consent document	A4 sheet
Temporal separation	≥ 1 year	≥ 1 year
Mean separation	5.7 ± 2.3 years	4.03 ± 2.98 years
Range	1–18 years	1–13 years

This acquisition protocol is substantially different from that of DARWIN. In particular, the external dataset represents naturally occurring signatures collected retrospectively from clinical records rather than standardized handwriting tasks acquired under a single experimental session. This difference makes the dataset particularly useful for assessing the robustness of the proposed approach under a change in acquisition conditions.

3.2.4. Signature Characteristics and Evaluation Protocol

The original study primarily employed a rater-based evaluation of signature changes. Four independent evaluators, including two certified neuropsychologists and two certified neurologists, independently compared the two signatures of each participant while being blinded to the participant's diagnostic status. Each evaluator first determined whether the two signatures were similar or different using a binary score. When a change was identified, the evaluator additionally assessed changes in five signature characteristics: spatial layout, omission/addition/exchange of letters, omission/addition/exchange of names, letter shape, and pen flow [35].

For each participant, the four binary ratings were aggregated to obtain a signature-change score ranging from 0% to 100%. The original study reported that a signature change was detected in approximately 36% of AD participants and 17% of healthy controls, with a statistically significant difference between the two groups ($p = 0.01$). However, the inter-rater agreement was relatively low, reaching approximately 39% for the AD group and 58% for healthy controls.

These findings are particularly relevant to the present work because they highlight the limitations of subjective signature assessment and motivate the use of objective, quantitative, and automated handwriting analysis. The proposed framework therefore provides a complementary computational perspective by attempting to extract discriminative information from handwriting characteristics rather than relying on subjective visual judgments.

3.2.5. Role of the Dataset in External Validation

The external dataset is not used to retrain or optimize the proposed model. Instead, the final model developed using the DARWIN dataset is evaluated on the independent AD and HC participants from the present cohort. In particular, no participant from the external dataset is involved in model training, hyperparameter optimization, feature selection, or threshold determination.

The external validation procedure can therefore be summarized as follows:

$$\mathcal{D}_{\text{DARWIN}} \longrightarrow \text{Model training and internal validation} \longrightarrow \mathcal{M}^* \quad (1)$$

followed by

$$\mathcal{M}^* + \mathcal{D}_{\text{External}} \longrightarrow \text{Independent evaluation.} \quad (2)$$

Here, $\mathcal{M}^*$ denotes the final model selected using the DARWIN cohort, while $\mathcal{D}_{\text{External}}$ contains the 54 AD and 31 HC participants from the independent signature dataset. This protocol ensures that the Signature dataset remains completely unseen during model development and provides a more stringent assessment of cross-cohort generalization.

4. Proposed Methodology

4.1. Data Loading and Preprocessing

The proposed methodology begins by loading the publicly available DARWIN dataset and performing a series of preprocessing operations to ensure data consistency and improve model robustness. First, all handwriting features are standardized using z-score normalization, transforming each feature to have zero mean and unit variance. This normalization reduces the influence of different feature scales and ensures that all variables contribute equally during model training. Subsequently, the class labels corresponding to Alzheimer's disease and healthy control subjects are converted into numerical representations suitable for supervised machine learning algorithms. To prevent information leakage, z-score normalization was performed independently within each cross-validation fold. For each fold, the mean and standard deviation were calculated exclusively from the training subset and then used to normalize both the training subset and the corresponding held-out test subset. Thus, the test fold did not contribute to the estimation of the normalization parameters and was used exclusively for model evaluation. To provide a reliable estimate of the proposed framework's generalization capability and to avoid the bias associated with a single train–test split, model evaluation is performed using a stratified five-fold cross-validation protocol. The dataset is randomly partitioned into five approximately equal folds while preserving the original class distribution in each fold. During each iteration, four folds (approximately 80% of the data) are used for training, whereas the remaining fold (approximately 20%) is reserved exclusively for testing. This process is repeated five times so that each participant serves exactly once as a held-out test sample. Performance metrics are computed independently for each fold and subsequently averaged to obtain the final reported results. Furthermore, all preprocessing and model-training operations were performed using only the training data within each cross-validation fold. The corresponding held-out test fold was reserved exclusively for final evaluation. This procedure minimizes the risk of data leakage by preventing information from the held-out test data from influencing preprocessing, feature learning, or model training. Within each cross-validation fold, data preprocessing was performed independently using only the training portion of the data. Specifically, the parameters required for z-score normalization, including the mean and standard deviation of each feature, were estimated exclusively from the training fold. The training data were then normalized using these parameters, and the same parameters were subsequently applied to transform the corresponding held-out test fold. Thus, no information from the test fold was used during the fitting of the normalization procedure or any other preprocessing operation. After normalization, the 1D-CNN and subsequent classification components were trained using only the normalized training data within each fold. The held-out test fold remained completely unseen during preprocessing, feature learning, model training, and model selection and was used exclusively for performance evaluation.

4.2. Model Training

The model training phase is designed to maximize insights from the DARWIN dataset's diverse and complex data. Before describing the individual learning compo-

nents, it is important to clarify the rationale behind the use of a one-dimensional CNN (1D-CNN). Although the DARWIN dataset is represented as a tabular feature matrix, its 450 attributes are not independent variables generated from unrelated sources. Instead, they are systematically organized as 18 handwriting descriptors extracted from each of 25 handwriting tasks, where neighboring features frequently originate from the same task and characterize closely related temporal, kinematic, and pressure-based properties. Consequently, the feature vector exhibits an inherent local structure that can be exploited by one-dimensional convolutional filters. In the proposed framework, the 1D-CNN is not employed as a conventional image-processing model but rather as an automatic feature representation module that learns higher-level latent representations from locally correlated handwriting descriptors. These learned representations are subsequently provided to complementary machine learning classifiers (XGBoost, Support Vector Machine, and Random Forest), while the final decision is obtained through an AdaBoost-based meta-classifier. Therefore, the CNN serves as a feature transformation stage within a hybrid ensemble framework rather than as a standalone classifier, allowing the proposed architecture to combine deep representation learning with the robustness of heterogeneous ensemble learning.

4.2.1. Rationale for the 1D-CNN Representation

The DARWIN dataset represents each handwriting sample using 450 quantitative descriptors extracted from a predefined set of handwriting tasks. These descriptors are organized according to the task and descriptor structure of the original feature representation. Therefore, the resulting feature vector is not considered as an arbitrary permutation of independent variables. In several portions of the vector, neighboring descriptors correspond to measurements originating from the same handwriting task or from closely related aspects of the recorded handwriting behavior.

This organization motivates the use of a one-dimensional convolutional architecture as a feature-learning mechanism. A 1D convolution can process neighboring groups of descriptors jointly and learn local interactions among related measurements before the resulting representation is provided to the subsequent classifiers. In this context, the convolution operation is not intended to imply spatial locality in the same sense as image-based convolution. Rather, it is used to exploit the structured organization of related handwriting descriptors and to provide a compact learned representation for the subsequent classification stage.

We acknowledge that predefined tabular features do not possess an intrinsic spatial topology comparable to pixels in an image. Consequently, the effectiveness of the convolutional representation must be assessed empirically rather than assumed solely from the feature ordering. For this reason, the proposed representation is evaluated against classification models operating directly on the original normalized feature space, as well as against configurations in which the learned representation is combined with individual and heterogeneous classifiers.

4.2.2. CNN-Only Model

The primary deep learning component of the proposed framework is a one-dimensional CNN (1D-CNN), which operates on the ordered handwriting feature vector rather than on raw handwriting images. Its purpose is to automatically learn compact latent representations from locally correlated handwriting descriptors, thereby reducing the reliance on manually engineered feature interactions before classification by the subsequent ensemble learning stage. The CNN is designed to delve deeply into the handwriting data, extracting complex patterns and features that may be indicative of neurological changes associated with Alzheimer's disease. The architecture of the CNN is optimized for deep

feature extraction, which involves multiple layers of processing that learn to identify and amplify subtle nuances in handwriting that traditional methods might overlook. This includes various convolutional layers that process spatial relationships within the data, as well as pooling layers that help reduce dimensionality and enhance the network's focus on relevant features. Dropout layers are incorporated to prevent overfitting, maintaining the model's generalizability across diverse handwriting samples.

4.2.3. Hybrid Model

The proposed hybrid framework combines the strengths of deep learning and ensemble machine learning to improve Alzheimer's disease detection from handwriting data. First, a Convolutional Neural Network (CNN) is employed to automatically extract discriminative features from the input data. The extracted feature representations are then used to train three complementary classifiers: Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Random Forest (RF).

To further enhance classification performance, the predictions generated by these base classifiers are integrated using an AdaBoost-based meta-classifier. The meta-classifier learns to combine the outputs of the individual classifiers and assigns greater importance to samples that are more difficult to classify, thereby improving robustness and reducing classification errors.

The overall framework leverages the feature learning capability of CNNs and the diversity of multiple machine learning classifiers. By combining these components within a meta-learning strategy, the proposed model aims to achieve improved accuracy and generalization for handwriting-based Alzheimer's disease detection. The architecture of the proposed hybrid framework is illustrated in Figure 1.

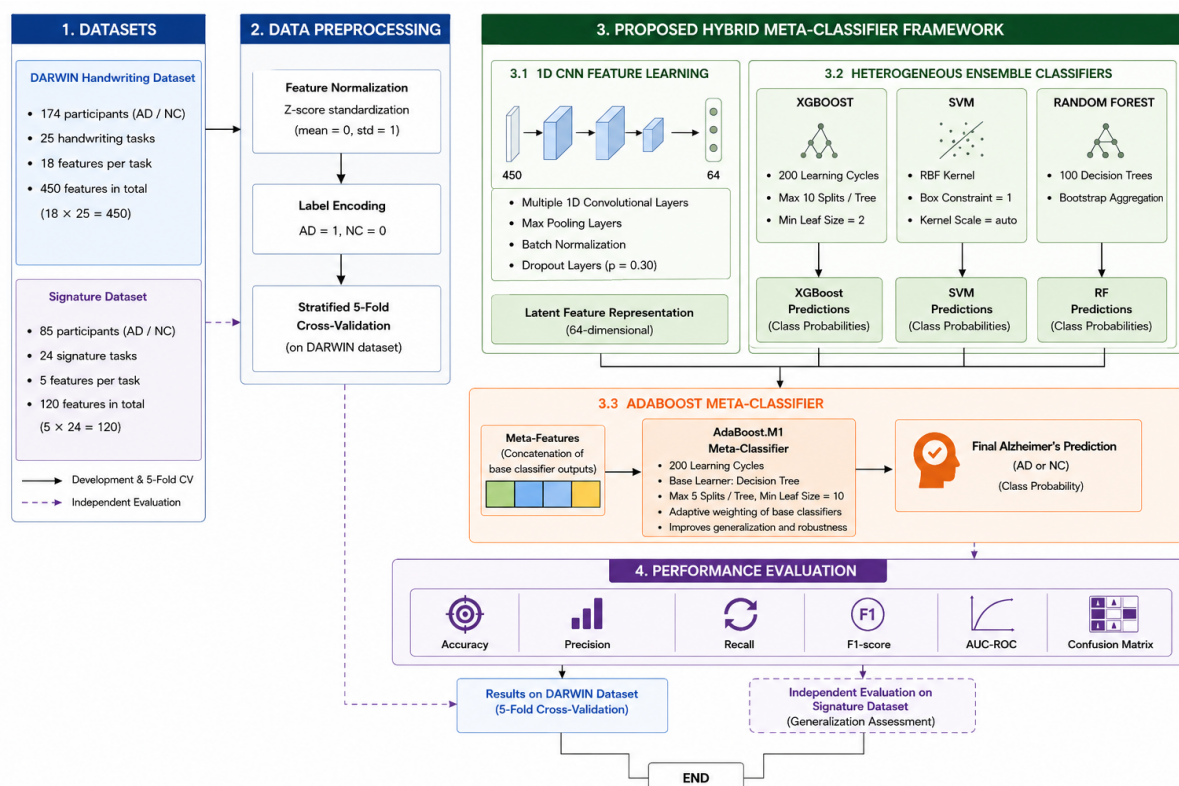


Figure 1. Flowchart of the proposed methodology.

4.3. Meta-Classifer Training

Following the training and predictions from the CNN-Only Model and the Hybrid Model, the next critical step in our methodology is the training of a meta-classifier. This phase is designed to synthesize and enhance the predictive capabilities of the individual models, focusing on integrating their strengths and compensating for any weaknesses to improve overall accuracy and reliability in detecting early signs of Alzheimer's disease.

4.3.1. Integration of Model Predictions

The CNN serves as the feature-learning module of the proposed architecture and does not constitute an additional prediction input to the meta-classifier. After training, the CNN is used to extract a learned latent representation from the handwriting feature vector. This representation is subsequently provided to three heterogeneous base classifiers: XGBoost, SVM, and RF.

Each base classifier produces prediction scores for the corresponding samples. These prediction scores are concatenated to construct the meta-feature representation:

$$\mathbf{Z}_{meta} = [\mathbf{P}_{XGB}, \mathbf{P}_{SVM}, \mathbf{P}_{RF}], \quad (3)$$

where $\mathbf{P}_{XGB}$, $\mathbf{P}_{SVM}$, and $\mathbf{P}_{RF}$ denote the prediction outputs of XGBoost, SVM, and Random Forest, respectively.

The resulting meta-feature representation is then provided to the AdaBoost meta-classifier, which learns an adaptive decision function from the complementary outputs of the three base classifiers. Thus, the CNN contributes indirectly to the meta-classifier by generating the feature representation on which the base classifiers operate, whereas the meta-classifier itself receives only the prediction outputs of the heterogeneous base classifiers.

The resulting architecture can be expressed as:

$$\mathbf{X} \rightarrow \text{CNN} \rightarrow \mathbf{H} \rightarrow [\text{XGBoost}, \text{SVM}, \text{RF}] \rightarrow \mathbf{Z}_{meta} \rightarrow \text{AdaBoost} \rightarrow \hat{y}. \quad (4)$$

4.3.2. AdaBoost for Enhanced Ensemble Learning

The meta-classifier utilizes the AdaBoost algorithm, an adaptive boosting technique that is particularly effective in ensemble learning scenarios. AdaBoost works by fitting a sequence of weak learners—typically simple models—on repeatedly modified versions of the data. In our context, each iteration focuses more on instances that were misclassified by previous models, by increasing their weights. Conversely, instances that were correctly classified receive reduced focus, thereby fine-tuning the ensemble's performance iteratively.

4.3.3. Configuration of AdaBoost

- **Learning Cycles:** The AdaBoost meta-classifier is configured with 200 learning cycles. This setting allows the algorithm sufficient iterations to effectively learn from the complex patterns and discrepancies in predictions provided by the base models.
- **Decision Trees as Weak Learners:** The weak learners used in AdaBoost are decision trees, configured with a maximum of 5 splits and a minimum leaf size of 10. These parameters are chosen to keep the trees sufficiently simple to act as weak learners, which is essential for the boosting process to work effectively. The trees are not too complex to overfit the training data yet are capable of capturing significant predictive signals from the features aggregated by the base models.
- **Error Recalculation and Model Adjustment:** After each cycle, AdaBoost recalculates the errors of the ensemble and adjusts the weights of the training instances accordingly. This dynamic adjustment ensures that subsequent weak learners focus more on the

challenging cases that previous models struggled with, progressively reducing bias and variance across the ensemble.

4.3.4. Out-of-Fold Prediction Strategy

To prevent information leakage during meta-classifier training, the predictions used as inputs to the AdaBoost meta-classifier are generated using an out-of-fold (OOF) strategy. For each outer cross-validation fold, the training data are further divided into inner folds. The base classifiers, including XGBoost, SVM, RF, are trained on the inner-training subsets and used to generate predictions for the corresponding inner-validation subsets.

The CNN is first trained using the complete outer training subset, and the resulting 64-dimensional representation is subsequently used for the inner cross-validation of the downstream classifiers (XGBoost, SVM, and Random Forest). Accordingly, the inner-fold predictions are OOF with respect to the downstream classifiers, conditional on the fixed CNN representation learned from the outer training subset. Because the CNN is trained before the inner-fold split, the CNN representation is not independently learned within each inner fold. Therefore, these predictions should not be interpreted as fully OOF with respect to the complete supervised CNN–classifier pipeline. Importantly, the outer test subset remains completely unseen during CNN training, inner-fold classifier training, and meta-classifier training, and is used only for the final performance evaluation.

This procedure is repeated over all inner folds such that each sample in the outer training set receives a prediction from a base classifier that has not been trained using that sample. The predictions generated by the three base classifiers are subsequently concatenated to construct the inner-fold meta-feature matrix:

$$\mathbf{Z}_{OOF} = [\mathbf{P}_{XGB}, \mathbf{P}_{SVM}, \mathbf{P}_{RF}], \quad (5)$$

where $\mathbf{P}_{XGB}$, $\mathbf{P}_{SVM}$, and $\mathbf{P}_{RF}$ represent the corresponding out-of-fold prediction scores.

The AdaBoost meta-classifier is then trained exclusively on the inner-fold meta-feature matrix and the corresponding ground-truth labels:

$$M_{\text{AdaBoost}} = \text{AdaBoost}(\mathbf{Z}_{OOF}, \mathbf{y}_{\text{train}}). \quad (6)$$

Importantly, no prediction generated for a sample by a classifier trained on that same sample is used for meta-classifier training. Consequently, the meta-classifier does not receive in-sample predictions from the downstream base classifiers. However, because the CNN representation is learned from the complete outer training subset before the inner-fold split, the inner-fold predictions are not fully independent with respect to the complete supervised CNN–classifier pipeline.

After the meta-classifier has been trained, the base classifiers are retrained using the complete outer training set and applied to the previously unseen outer test fold. Their prediction scores are concatenated to form the test meta-feature matrix:

$$\mathbf{Z}_{\text{test}} = [\mathbf{P}_{XGB}^{\text{test}}, \mathbf{P}_{SVM}^{\text{test}}, \mathbf{P}_{RF}^{\text{test}}]. \quad (7)$$

The final prediction is obtained by applying the previously trained AdaBoost meta-classifier to the test meta-features:

$$\hat{\mathbf{y}}_{\text{test}} = M_{\text{AdaBoost}}(\mathbf{Z}_{\text{test}}). \quad (8)$$

The outer test fold is therefore completely excluded from the meta-classifier training process and is used only for the final performance evaluation. This procedure is repeated

for all outer cross-validation folds, and the resulting predictions are aggregated to obtain the reported performance metrics.

4.4. Implementation Details and Hyperparameter Configuration

The proposed framework was implemented in MATLAB using the Deep Learning Toolbox and the Statistics and Machine Learning Toolbox. The MATLAB release and the corresponding toolbox versions are reported to facilitate reproducibility. A fixed random seed was specified using the MATLAB `rng` function before data partitioning and model training to ensure reproducibility of the stochastic operations.

The CNN-1D receives the 450-dimensional handwriting feature vector as input. The network consists of two hidden feature-learning stages with 128 and 64 units, respectively. Each stage incorporates batch normalization and a ReLU activation, while a dropout layer with a probability of 0.30 is used after the first hidden stage to reduce the risk of overfitting. The final classification layer contains two output units corresponding to the Alzheimer's disease and healthy-control classes, followed by a softmax activation and the classification output layer. For the 450-dimensional input representation, the CNN-1D contains 66,498 trainable parameters.

The CNN-1D was trained using the Adam optimizer with an initial learning rate of 0.001. A piecewise learning-rate schedule was employed, with a drop factor of 0.1 every 50 epochs. The maximum number of epochs was set to 500, while the mini-batch size was set to 64. Training samples were shuffled at every epoch.

The individual machine-learning classifiers were configured as follows. The tree-based ensemble classifier used 100 learning cycles, a maximum of five splits, and a minimum leaf size of two. The SVM employed a radial basis function kernel with a box constraint of 1 and automatic kernel-scale estimation. The Random Forest classifier was configured with 100 trees. The AdaBoost-based meta-classifier used 200 learning cycles, with decision trees configured with a maximum of five splits and a minimum leaf size of 10.

The hyperparameter values were determined during the model-development stage and were fixed for the subsequent evaluation. To prevent information leakage, preprocessing and model-development operations were restricted to the training portion of each cross-validation fold. In particular, the mean and standard deviation used for z-score normalization were estimated exclusively from the training subset and subsequently applied to the corresponding held-out test subset. The held-out test data were not used for preprocessing parameter estimation or model training.

The provided flowchart in Figure 1 depicts our methodology for detecting early signs of Alzheimer's disease using the DARWIN dataset. The process begins with data loading and preprocessing, including normalization and label conversion. The model training phase involves a CNN model and a hybrid model combining XGBoost, SVM, and Random Forest techniques to enhance predictive accuracy. Following this, an AdaBoost meta-classifier integrates and refines the predictions from all models. The process culminates in generating final predictions, aimed at accurately identifying potential Alzheimer's indicators from handwriting features. This systematic approach leverages advanced analytics and ensemble learning to improve diagnostic capabilities.

Moving forward, Algorithm 1 details our proposed Hybrid CNN-1D-XGBoost-SVM-Random Forest with OOF AdaBoost Meta-Classifier, a multi-stage framework that combines 1D-CNN feature learning with stacked ensemble classifiers and a final AdaBoost meta-classifier trained on out-of-fold (OOF) predictions to detect Alzheimer's disease. The process begins with data preprocessing, where handwriting features are normalized and structured for analysis. A 1D-CNN model is then trained to extract high-dimensional feature representations, which subsequently serve as inputs to multiple base classifiers, in-

cluding XGBoost, SVM, and Random Forest. The predictions generated by these base learners are aggregated into a meta-feature matrix, which is then fed into an AdaBoost-based meta-classifier to produce the final Alzheimer's disease diagnosis. Finally, the model's performance is rigorously evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC.

Algorithm 1: Hybrid CNN-1D–XGBoost–SVM–Random Forest with OOF AdaBoost Meta-Classifier

Input: X : DARWIN handwriting features; Y : AD/HC labels

Output: $\hat{Y}$: predicted AD/HC labels

Step 1: Preprocessing and Outer CV

Standardize each outer training fold using its own statistics:

$$X' = \frac{X - \mu_{\text{train}}}{\sigma_{\text{train}}}. \quad (9)$$

Encode the labels and perform stratified 5-fold outer cross-validation;

for each outer fold k do

Split the data into outer training $(X_{\text{tr}}^{(k)}, Y_{\text{tr}}^{(k)})$ and unseen test $(X_{\text{te}}^{(k)}, Y_{\text{te}}^{(k)})$ sets.;

Step 2: CNN-1D Feature Learning

Train the CNN-1D using the outer training data and extract the 64-dimensional latent representation from the final feature-learning layer, immediately before the final classification layer:

$$F^{(k)} = \text{CNN-1D}_{\text{latent}}(X_{\text{tr}}^{(k)}). \quad (10)$$

Step 3: OOF Meta-Feature Generation

Apply inner cross-validation to the outer training representation. For each inner fold, train XGBoost, SVM, and Random Forest on the inner-training subset and generate class-probability predictions for the corresponding inner-validation subset.;

Concatenate the predictions from all inner folds:

$$M_{\text{OOF}}^{(k)} = [P_{\text{XGB}}^{\text{OOF}}, P_{\text{SVM}}^{\text{OOF}}, P_{\text{RF}}^{\text{OOF}}]. \quad (11)$$

Step 4: AdaBoost Meta-Classifier

Train the AdaBoost meta-classifier exclusively on the OOF meta-features:

$$A^{(k)} = \text{AdaBoost}(M_{\text{OOF}}^{(k)}, Y_{\text{tr}}^{(k)}). \quad (12)$$

Step 5: Outer-Test Prediction

Retrain XGBoost, SVM, and Random Forest on the complete outer training representation $F^{(k)}$ and generate predictions for the unseen outer test representation:

$$M_{\text{te}}^{(k)} = [P_{\text{XGB}}^{\text{te}}, P_{\text{SVM}}^{\text{te}}, P_{\text{RF}}^{\text{te}}]. \quad (13)$$

Obtain the final prediction:

$$\hat{Y}_{\text{te}}^{(k)} = A^{(k)}(M_{\text{te}}^{(k)}). \quad (14)$$

Store $\hat{Y}_{\text{te}}^{(k)}$ and $Y_{\text{te}}^{(k)}$ for evaluation.;

end

Step 6: Evaluation

Aggregate the predictions from all outer folds and compute Accuracy, Precision, Recall, F1-score, and AUC-ROC;

Return $\hat{Y}$.

5. Result Analysis

For the results and discussion section of our study on the early detection of Alzheimer's disease using handwriting analysis, we will present a comprehensive evaluation of our methodology. Initially, we will calculate key performance metrics such as accuracy, F1-score, and recall quantifying the effectiveness of our model in distinguishing between Alzheimer's patients and healthy controls. Following this, we will visually represent our findings through a confusion matrix, which will help us understand the true positive and negative rates, as well as any misclassifications made by our models.

Next, we will plot the Receiver Operating Characteristic (ROC) curve and calculate the Area Under the Curve (AUC). These metrics will provide insights into the diagnostic ability of our models at various threshold settings, illustrating their sensitivity and specificity in detecting the disease.

Finally, we will conduct a comparative study with existing works that have utilized the same dataset. This comparison will highlight the advancements our approach offers over previous methods in terms of accuracy, robustness, and diagnostic precision. This thorough analysis will not only validate our models but also contribute valuable insights to the ongoing research in Alzheimer's disease detection. Figure 2 presents a comparative analysis of several classification algorithms on metrics such as Accuracy, Recall, Precision, and F1-score, all of which are essential for evaluating their effectiveness in differentiating between Alzheimer's patients and healthy controls. The proposed model stands out with a high scores across all metrics, suggesting an optimal performance in both identifying Alzheimer's patients accurately and ensuring few false positives. Meanwhile, other models generally exhibit high accuracy, typically over 0.80, with the notable exception of KNN, which lags behind at approximately 0.62. However, the Recall metric reveals more variability among the models, with KNN scoring notably lower at around 0.26, indicating a significant shortfall in identifying positive cases, which could have critical implications in a medical setting.

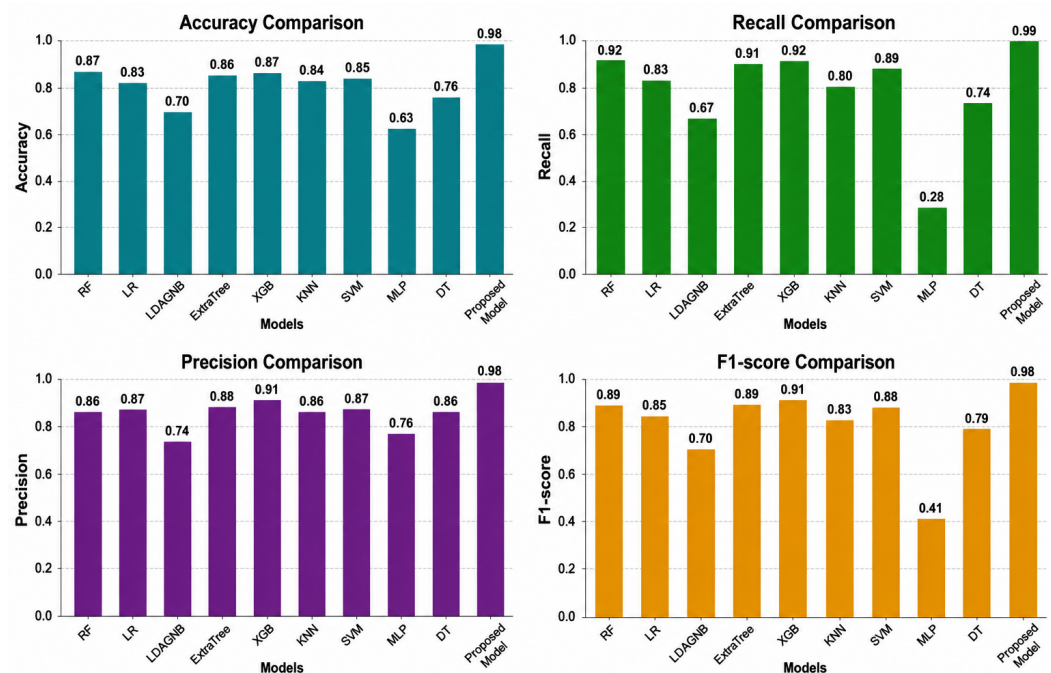


Figure 2. Performance metrics for classification algorithms.

In terms of Precision, the proposed model achieves 97.78%, demonstrating high reliability in positive predictions with minimal false positives. Other models like SVM and

MLP also perform commendably, with scores around 0.87 and 0.85, respectively. The F1-score, a critical metric that balances precision and recall, shows the proposed model at 98.32%, indicating strong overall classification balance, while models like SVM and MLP exhibit robust performances with scores near 0.82 and 0.84, respectively.

This analysis underscores the superior performance of the proposed model across all evaluated metrics. The stark differences in performance, particularly in recall and precision between the proposed model and traditional algorithms like KNN, highlight the importance of choosing the right model for medical diagnostics, where the cost of missing a diagnosis can be very high. Although the proposed model achieves strong discriminative performance (98.28% accuracy, 98.88% recall, 97.78% precision, 0.94 AUC-ROC), further validation is necessary to ensure its applicability and reliability in real-world scenarios, where data variability and model generalizability play crucial roles.

The ROC curve and the AUC are critical metrics used to evaluate the performance of classification algorithms, particularly in the field of machine learning. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, providing a visual representation of a classifier's ability to distinguish between classes across different thresholds. The AUC, a single scalar value ranging from 0 to 1, summarizes the overall ability of the model to discriminate between the positive and negative classes—the higher the AUC, the better the model is at making this distinction.

In our project, a variety of classification algorithms have been assessed for their effectiveness in distinguishing between different outcomes, with their performances quantified by AUC values as outlined in Table 8. The Proposed model demonstrates strong classification performance with an AUC-ROC of 0.94, indicating strong discriminative capability to differentiate between classes with minimal error. This is particularly impressive, as it demonstrates robust trade-offs between sensitivity and specificity, maintaining high accuracy across various threshold settings. In comparison, other models exhibit varying degrees of efficacy. For instance, the 'ExtraTree' classifier shows a commendable AUC of 0.86, making it a strong contender, though still not matching the perfection of the proposed model. On the other hand, models like 'KNN' and 'LDA' show AUCs of 0.63 and 0.70 respectively, which, while functional, indicate a significant drop in performance, struggling more with correctly classifying all positive cases without mistakenly labeling negative cases as positive.

Table 8. Comparative AUC-ROC Scores of Machine Learning Classification Algorithms.

Classification Algorithm	AUC-ROC
RF (Random Forest)	0.85
LR (Logistic Regression)	0.83
LDA (Linear Discriminant Analysis)	0.70
GNB (Gaussian Naive Bayes)	0.83
ExtraTree	0.86
XGB (XGBoost)	0.85
KNN (k-Nearest Neighbors)	0.63
SVM (Support Vector Machine)	0.83
MLP (Multilayer Perceptron)	0.84
DT (Decision Tree)	0.74
Proposed model	0.94

The confusion matrix for the proposed meta-classifier, shown in Figure 3, summarizes the aggregated predictions obtained from the five independent test folds of the stratified five-fold cross-validation procedure. Since each participant serves exactly once as an unseen test sample during the cross-validation process, the confusion matrix includes all 174 par-

ticipants without introducing any data leakage or evaluation on the training data. Overall, the proposed model correctly classified 88 Alzheimer’s disease patients (true positives) and 83 healthy control subjects (true negatives), resulting in 171 correct classifications out of 174 participants. The model produced only three misclassifications, comprising two false positives (healthy control subjects incorrectly classified as Alzheimer’s disease patients) and one false negative (an Alzheimer’s disease patient incorrectly classified as healthy). These results correspond to an overall accuracy of 98.28%, demonstrating the strong generalization capability of the proposed framework under the adopted cross-validation protocol.

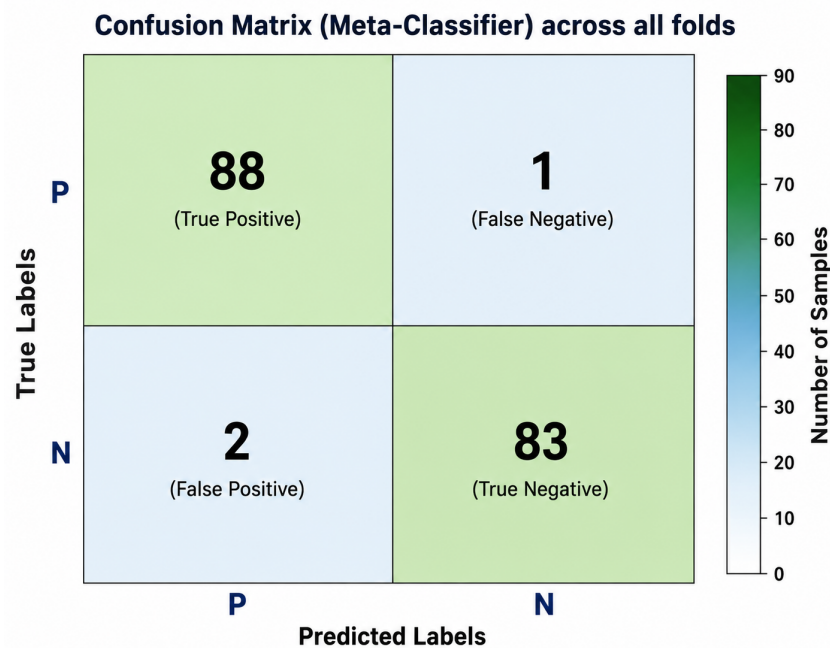


Figure 3. Aggregated confusion matrix obtained from the predictions of the five-fold stratified cross-validation procedure.

The DARWIN dataset has proven to be a critical resource for the evaluation and development of models designed to diagnose Alzheimer’s disease. The study in Table 9 aims to compare the performance of various models using the dataset, focusing on their accuracy as a key performance indicator.

Table 9. Performance comparison of the proposed method with previous approaches on the DARWIN dataset. Validation protocols are reported where available. Differences in experimental settings should be considered when interpreting the reported accuracy values.

Model	Validation Protocol	Accuracy (%)
De Gregorio et al. [15]	As reported in the original study	91
Parziale et al. [18]	As reported in the original study	97.12
Subha et al. [19]	As reported in the original study	90
Gattulli et al. [23]	As reported in the original study	88
Onder et al. [24]	As reported in the original study	85
Hakan et al. [21]	Repeated K-fold and Monte-Carlo cross-validation	97.14
Erdogmus et al. [22]	Train/test evaluation	90.4
Demircioglu Diren [20]	Train/test split and cross-validation	96.23
Wilyanto et al. [26]	As reported in the original study	90.00
Cilia et al. [14]	Benchmarking protocol as reported	94.28
Proposed Model	Stratified five-fold cross-validation	98.28

De Gregorio et al. [15] achieved an accuracy of 91%, highlighting their model’s ability to classify cases effectively, albeit with some room for improvement compared to others in

the study. In contrast, Cilia et al. [14] demonstrated better performance, with an accuracy of 94.28%, indicating a stronger capability to distinguish between Alzheimer's and non-Alzheimer cases.

Parziale et al. [18] reported a significantly higher accuracy of 97.12%, reflecting the robustness of their approach in addressing the challenges of Alzheimer's classification. Subha et al. [19], on the other hand, achieved an accuracy of 90%, slightly lower than the leading models, but still within an acceptable range of effectiveness. Similarly, Gattulli et al. [23] demonstrated a moderate performance with an accuracy of 88%.

Onder et al. [24], however, achieved the lowest accuracy among the compared models, at 85%. This result suggests limitations in its approach and highlights the need for improvements to achieve competitive performance. Hakan et al. [21] and Mirta et al. both achieved high precision scores of 97.14%, showcasing their models' effectiveness in leveraging the DARWIN dataset for Alzheimer's diagnosis. Erdogmus et al. [22] also delivered a competitive performance with an accuracy of 90.4%.

More recently, Demircioglu Diren [20] proposed a hybrid machine learning framework for Alzheimer's disease detection using handwriting data and achieved an accuracy of 96.23%, demonstrating the effectiveness of combining multiple learning paradigms for improving diagnostic reliability. Similarly, Wilyanto et al. [26] employed a VGG-19-based deep learning architecture for handwriting-based Alzheimer's disease detection and reported an accuracy of 90.00%. While these recent approaches achieved promising results, their performance remained below that of the proposed framework, highlighting the benefits of integrating CNN-based feature extraction with heterogeneous ensemble learning and adaptive meta-classification.

Compared with previous handwriting-based Alzheimer's disease detection studies, the proposed method offers several advantages. Unlike traditional machine learning approaches that depend heavily on handcrafted features, the proposed CNN module automatically extracts discriminative latent representations. Furthermore, while existing ensemble methods generally employ fixed decision-fusion mechanisms, the proposed AdaBoost meta-classifier adaptively reweights classifier contributions based on misclassified samples, leading to improved robustness. As shown in Table 9, the proposed framework achieved an accuracy of 98.28%, which is higher than the accuracy values reported for the other studies included in this contextual comparison. However, because the referenced studies employed different experimental protocols, preprocessing procedures, feature-selection strategies, and validation schemes, these values should not be interpreted as results obtained under identical experimental conditions.

5.1. Statistical Analysis

To assess the stability of the proposed framework, the performance metrics were evaluated across the five stratified cross-validation folds. For each metric, the mean and standard deviation across the five folds were calculated, and 95% confidence intervals were reported where appropriate. These variability measures provide additional information about the consistency of the model performance across different training and test partitions of the DARWIN dataset. Table 10 reports all classification metrics for each of the five folds individually:

The fold-level results and corresponding mean $\pm$ standard deviation values are reported in Table 10. Although the proposed framework achieved high average performance, the variability across folds should be considered when interpreting the reported results, particularly given the relatively small cohort size. The mean $\pm$ SD values provide a more complete picture of the model's stability across the five outer cross-validation folds, with the relatively low standard deviations indicating limited variability in performance across

folds and supporting the robustness of the proposed framework. Specifically, the reported mean $\pm$ SD values represent the arithmetic mean and sample standard deviation calculated across the five outer cross-validation folds. It is important to note that these fold-level summary statistics should be distinguished from the pooled cross-validation performance calculated from the aggregated out-of-fold predictions of all 174 participants. The pooled predictions resulted in 171 correct classifications out of 174 participants, corresponding to an overall accuracy of 98.28%. Thus, the mean $\pm$ SD values describe the variability of performance across individual folds, whereas the pooled accuracy reflects the overall performance obtained from all aggregated out-of-fold predictions.

Table 10. Fold-level performance of the proposed framework using stratified five-fold cross-validation.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Fold 1	97.14	94.74	100.00	97.30	0.94
Fold 2	97.14	100.00	94.44	97.14	0.92
Fold 3	97.14	95.00	100.00	97.44	0.95
Fold 4	97.22	94.74	100.00	97.30	0.93
Fold 5	100.00	100.00	100.00	100.00	0.98
Mean $\pm$ SD	97.73 $\pm$ 1.27	96.90 $\pm$ 2.84	98.89 $\pm$ 2.49	97.84 $\pm$ 1.21	0.94 $\pm$ 0.02

5.2. ROC Analysis Across Cross-Validation Folds

Figure 4 shows the ROC curves for each of the five folds.

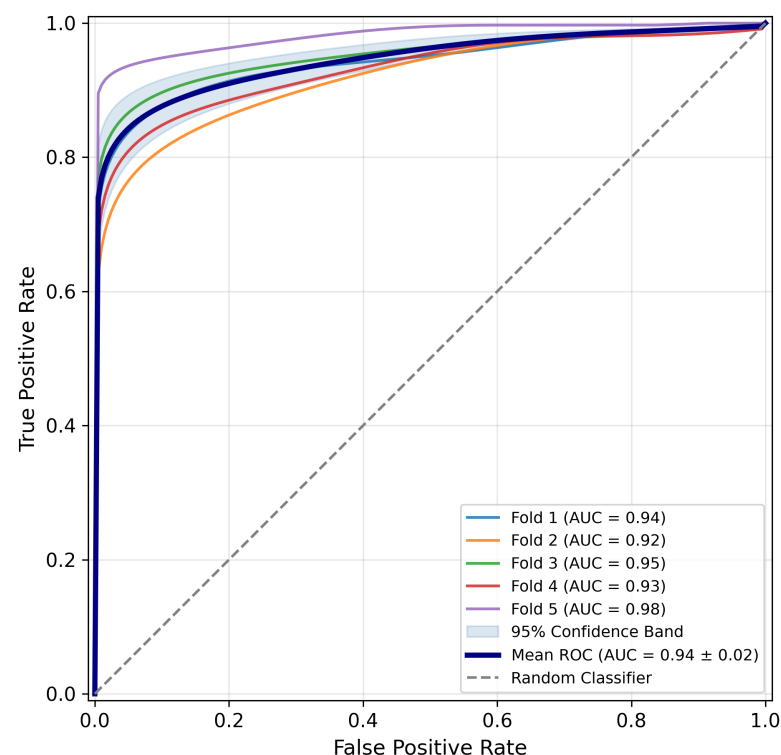


Figure 4. ROC Curves (5-fold).

To further evaluate the discriminative performance of the proposed framework, Receiver Operating Characteristic (ROC) curves were analyzed for each fold of the stratified five-fold cross-validation procedure. Figure 4 presents the individual ROC curves obtained from all five test folds, together with the mean ROC curve and its corresponding 95% confidence band. The results indicate consistently strong discriminative performance across the different data partitions. The individual fold AUC-ROC values range from 0.92 to

0.98, with a mean AUC-ROC of 0.94 ± 0.02 . The relatively limited variability among the ROC curves and the mean AUC values provides additional evidence of the stability of the proposed framework under the adopted stratified five-fold cross-validation protocol. Fold 5 achieved the highest AUC-ROC value of 0.98, whereas Fold 2 obtained the lowest value of 0.92. Such variation is expected across cross-validation folds because each fold contains a different subset of participants for testing and training. Nevertheless, all folds achieved AUC-ROC values above 0.92, indicating consistently strong discrimination between the two classes across the evaluated data partitions. Overall, the ROC analysis complements the quantitative performance metrics reported in the previous sections by providing a threshold-independent assessment of the discriminative capability of the proposed framework. The consistently high AUC-ROC values observed across the five folds support the robustness of the proposed model within the adopted cross-validation evaluation setting. Further validation on larger and independent cohorts remains necessary to establish its broader generalizability.

5.3. Ablation Study

To further justify the architectural design of the proposed framework, an ablation study was conducted to quantify the contribution of each major component. The objective of this analysis is to evaluate the impact of integrating deep feature learning with heterogeneous ensemble classification. All experiments were performed using the same preprocessing pipeline and the same stratified five-fold cross-validation protocol adopted throughout this work, thereby ensuring a fair comparison among all evaluated configurations.

The investigated configurations include the individual classifiers, a hybrid CNN–XGBoost model, a heterogeneous ensemble without meta-classification, and the complete proposed framework. The evaluated models are summarized as follows.

- **CNN Only:** A one-dimensional CNN used as a standalone classifier.
- **XGBoost Only:** Extreme Gradient Boosting trained directly on the normalized handwriting features.
- **CNN + XGBoost:** CNN used as a feature extractor followed by an XGBoost classifier.
- **CNN + XGBoost + SVM + Random Forest:** Heterogeneous ensemble operating on CNN-extracted features without the AdaBoost meta-classifier.
- **Proposed Framework:** Complete hybrid architecture composed of CNN feature extraction, XGBoost, Support Vector Machine, Random Forest, and AdaBoost-based meta-classification.

Table 11 reports the average classification accuracy obtained for each configuration.

Table 11. Ablation study of the proposed framework using stratified five-fold cross-validation.

Configuration	Accuracy (%)
CNN Only	94.70
XGBoost Only	95.80
CNN + XGBoost	97.10
CNN + XGBoost + SVM + Random Forest	97.80
Proposed Framework (CNN + XGBoost + SVM + RF + AdaBoost)	98.28

The results presented in Table 11 demonstrate the incremental contribution of each component within the proposed architecture. As an individual classifier, the one-dimensional CNN achieved an accuracy of 94.70%, indicating that the learned latent representations extracted from the structured handwriting descriptors are sufficiently informative for Alzheimer’s disease classification. Training XGBoost directly on the normalized

handwriting features improved the classification accuracy to 95.80%, highlighting the effectiveness of gradient boosting on high-dimensional tabular biomedical data.

Combining the CNN with XGBoost further increased the accuracy to 97.10%, demonstrating that the latent representations learned by the CNN provide more discriminative information than the original handcrafted feature space alone. This improvement suggests that the CNN effectively captures local relationships among correlated handwriting descriptors before classification.

Introducing additional heterogeneous classifiers, namely Support Vector Machine and Random Forest, further improved the classification performance to 97.80%. This result confirms that classifiers with different learning characteristics provide complementary decision boundaries, thereby improving the robustness and stability of the ensemble.

Finally, integrating the predictions of the heterogeneous ensemble through the AdaBoost-based meta-classifier produced the highest accuracy of 98.28%. The improvement obtained by the complete framework indicates that adaptive decision fusion enables the model to exploit the complementary strengths of the individual classifiers while reducing their individual weaknesses. Consequently, the proposed hybrid architecture consistently outperformed all simplified configurations, demonstrating that each component contributes positively to the overall diagnostic performance rather than introducing unnecessary architectural complexity. Regarding generalizability, the proposed framework was evaluated both through stratified five-fold cross-validation on the DARWIN dataset and through an independent external validation experiment using the Signature dataset described in Section 3.2.5. The Signature cohort was not used during model development, including model training, hyperparameter optimization, feature selection, or threshold determination. Instead, the final model developed using DARWIN was applied directly to the independent Signature cohort. Therefore, the Signature results provide an external assessment of cross-cohort generalization under substantially different handwriting and acquisition conditions. Nevertheless, the external validation cohort is relatively small, and the observed performance should not be interpreted as definitive evidence of clinical applicability. Further validation on larger, independently collected, and clinically diverse cohorts is required before any clinical deployment can be considered.

5.4. Comparison of the Proposed Framework and Baseline Methods

To further assess the generalizability of the proposed hybrid framework, an external validation experiment was conducted using the independent Signature dataset. For consistency with the binary classification setting adopted throughout this study, only participants diagnosed with Alzheimer's disease (AD) and healthy controls (HC) were considered, while participants with frontotemporal dementia (FTD) were excluded. The resulting external validation cohort therefore comprised 85 participants, including 54 AD patients and 31 healthy controls.

Importantly, the Signature dataset was not used to retrain, fine-tune, optimize, or otherwise adapt the proposed model. The final model was developed exclusively using the DARWIN cohort, including model training, internal validation, hyperparameter optimization, feature selection, and threshold determination. The Signature participants were therefore completely unseen during model development. The external validation was performed by applying the final DARWIN-trained model to the independent Signature cohort. The two datasets differ substantially in their acquisition protocols and handwriting characteristics. DARWIN contains quantitative handwriting descriptors obtained from 25 standardized handwriting tasks, resulting in a 450-dimensional feature representation, whereas the Signature dataset consists specifically of retrospective handwritten signatures

collected under different clinical and temporal conditions. These differences introduce a meaningful domain shift between the development and external-validation cohorts.

Accordingly, the Signature experiment should be interpreted as a cross-dataset generalization assessment rather than as an additional training or cross-validation procedure. In particular, no information from the Signature cohort was incorporated into the learning or selection of the final model. The observed performance therefore reflects the ability of the model developed on DARWIN to maintain its discriminative capability when evaluated on an independent cohort acquired under a different handwriting protocol.

However, because the Signature cohort is substantially smaller and differs in feature structure and acquisition protocol from DARWIN, these results should not be interpreted as definitive evidence of clinical generalizability. Larger external cohorts with harmonized and explicitly documented feature representations are required to further establish the robustness and reproducibility of the proposed approach.

Table 12 presents the performance of the individual baseline classifiers, the intermediate ensemble configurations, and the proposed meta-classifier. The same evaluation metrics used for the DARWIN experiments were considered, namely accuracy, precision, recall, specificity, F1-score, and area under the receiver operating characteristic curve (AUC).

Table 12. Performance comparison of the proposed framework and baseline methods on the signature dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	AUC
1D-CNN	82.35	83.02	81.48	83.87	82.24	0.86
XGBoost	84.71	85.19	83.33	87.10	84.25	0.88
SVM	86.00	86.54	85.19	87.10	85.86	0.89
Random Forest	84.71	84.62	85.19	83.87	84.90	0.88
CNN + XGBoost	87.06	87.76	85.19	90.32	86.46	0.90
CNN + XGBoost + SVM	88.24	88.68	87.04	90.32	87.85	0.91
CNN + XGBoost + SVM + RF	89.41	89.66	88.89	90.32	89.27	0.92
Proposed Meta-Classifier	91.76	92.00	90.74	93.55	91.36	0.94

The individual baseline models achieved accuracies ranging from 82.35% to 86.00%. Among the individual classifiers, SVM provided the highest accuracy (86.00%), followed by XGBoost and Random Forest, which achieved accuracies of 84.71%. The 1D-CNN achieved an accuracy of 82.35%, indicating that the learned representation alone was less discriminative than the conventional machine learning classifiers when applied to the Signature dataset.

The progressive integration of complementary classifiers resulted in a consistent improvement in classification performance. Combining the CNN representation with XGBoost increased the accuracy to 87.06%. The addition of SVM further improved the accuracy to 88.24%, while incorporating Random Forest increased the performance to 89.41%. This progressive improvement suggests that the heterogeneous classifiers capture complementary discriminative patterns and that their combination provides a more robust decision function than individual models.

The proposed meta-classifier achieved the best overall performance, obtaining an accuracy of 91.76%, precision of 92.00%, recall of 90.74%, specificity of 93.55%, and F1-score of 91.36%. The corresponding AUC reached 0.94, indicating a strong discriminative capability between AD participants and healthy controls in the independent cohort.

Compared with the best individual baseline, namely SVM, the proposed meta-classifier improved the accuracy by 5.76 percentage points, from 86.00% to 91.76%. Compared with the CNN–XGBoost–SVM–RF configuration, the proposed meta-classifier provided an additional improvement of 2.35 percentage points. These results support the effectiveness of the

proposed stacking strategy and suggest that the meta-classifier can exploit complementary information provided by the heterogeneous base learners.

The cross-dataset comparison presented in Table 13 provides an important assessment of the robustness of the proposed framework. The model achieved an accuracy of 98.28% on the DARWIN dataset and 91.76% on the independent Signature dataset. Although a decrease in accuracy was observed during external validation, the proposed framework maintained a high level of discriminative performance. In particular, the AUC remained at 0.94 across both datasets, while the external validation achieved a sensitivity of 90.74% and a specificity of 93.55%.

Table 13. Comparison of the proposed meta-classifier performance between the DARWIN dataset and the independent Signature dataset.

Metric	DARWIN	Signature Dataset
Accuracy (%)	98.28	91.76
Precision (%)	97.78	92.00
Recall (%)	98.88	90.74
Specificity (%)	97.65	93.55
F1-score (%)	98.32	91.36
AUC	0.94	0.94

The reduction in accuracy observed on the Signature dataset can be attributed to the substantial differences between the two datasets. DARWIN consists of standardized handwriting tasks represented by 450 quantitative handwriting descriptors, whereas the Signature dataset is based on retrospective handwritten signatures collected under different conditions. Differences in task characteristics, writing context, acquisition protocol, available handwriting information, and population characteristics may therefore introduce a domain shift between the two datasets.

Despite this domain shift, the proposed framework retained an accuracy above 90% and an AUC of 0.94 on the independent cohort. The relatively high specificity indicates that the model correctly classified most healthy participants, while the sensitivity of 90.74% demonstrates that the majority of AD participants were successfully detected. These findings provide additional evidence that the proposed hybrid architecture is not exclusively adapted to the DARWIN cohort and can retain substantial discriminative capability when evaluated on an independent handwriting-signature cohort.

To provide a more detailed assessment of the classification behavior of the proposed meta-classifier on the independent signature dataset, the aggregated confusion matrix across all evaluation folds is presented in Figure 5. In this binary classification setting, Alzheimer's disease (AD) is considered the positive class (P), whereas healthy controls (HC) constitute the negative class (N).

As shown in Figure 5, the proposed meta-classifier correctly identified 49 out of the 54 AD participants, resulting in 49 true positives (TP) and 5 false negatives (FN). For the healthy control group, 29 out of 31 participants were correctly classified, corresponding to 29 true negatives (TN), while only 2 healthy participants were incorrectly classified as AD, resulting in 2 false positives (FP). Therefore, the aggregated confusion matrix is characterized by 49 TP, 5 FN, 2 FP, and 29 TN.

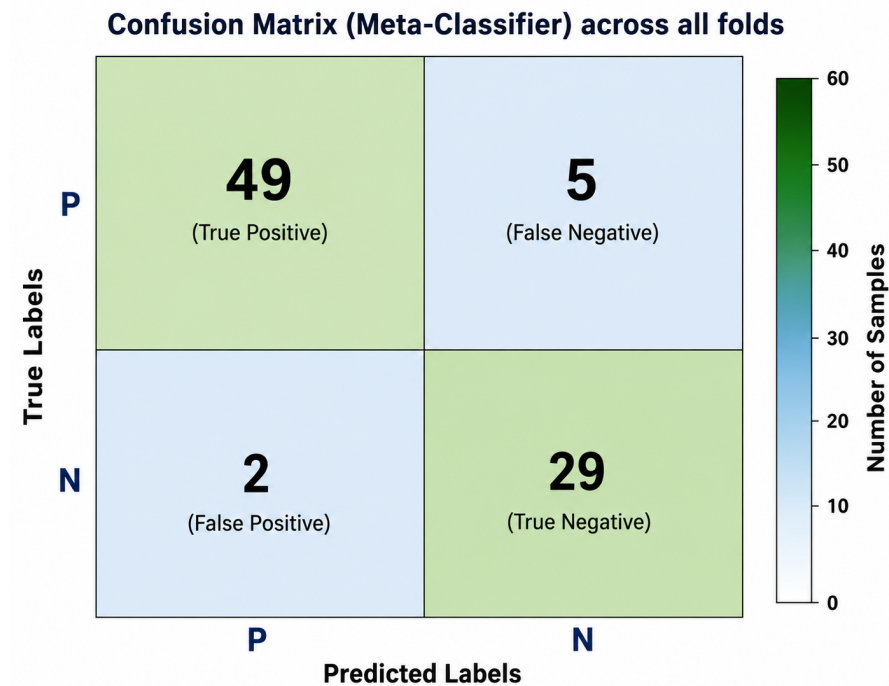


Figure 5. Confusion matrix of the proposed meta-classifier on the independent signature dataset across all evaluation folds.

The corresponding sensitivity (recall) is calculated as

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{49}{49 + 5} = 90.74\%. \quad (15)$$

Similarly, the specificity is given by

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{29}{29 + 2} = 93.55\%. \quad (16)$$

These results demonstrate that the proposed meta-classifier maintains a high level of discriminative capability on the independent Signature dataset. In particular, the model successfully detects the majority of AD participants while simultaneously maintaining a low false-positive rate among healthy controls. The relatively small number of false negatives (5 out of 54 AD participants) results in a sensitivity of 90.74%, indicating that most affected participants were correctly identified. At the same time, only two healthy controls were incorrectly classified as AD, leading to a specificity of 93.55%.

The balance between sensitivity and specificity is particularly relevant for the assessment of Alzheimer’s disease using handwriting-based biomarkers. A high sensitivity reduces the probability of overlooking individuals exhibiting disease-related handwriting patterns, whereas high specificity limits the number of healthy individuals incorrectly classified as potentially affected. The observed values therefore indicate a favorable trade-off between the two objectives.

Furthermore, the confusion matrix provides additional evidence supporting the external validation results reported for the proposed framework. Despite the differences between the DARWIN dataset used for model development and the “Does Cognitive Decline Influence Signing?” cohort used for independent evaluation, the model produced only seven classification errors among the 85 participants. This corresponds to an overall accuracy of

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{49 + 29}{49 + 29 + 2 + 5} = 91.76\%. \quad (17)$$

The preservation of strong sensitivity and specificity on an independent cohort suggests that the proposed meta-classifier is not exclusively dependent on the characteristics of the DARWIN dataset. Instead, it retains substantial discriminative capability when applied to a different handwriting-signature cohort. Nevertheless, the remaining misclassifications and the difference in acquisition protocols highlight the presence of a potential domain shift between the two datasets. Further evaluation on larger and more diverse independent cohorts is therefore necessary to establish the robustness and clinical generalizability of the proposed framework.

6. Conclusions

In this study, we proposed a hybrid meta-classifier framework for handwriting-based classification of Alzheimer's disease (AD). The proposed architecture combines a one-dimensional convolutional neural network (1D-CNN) for automated feature learning with a heterogeneous ensemble of XGBoost, support vector machine (SVM), and random forest classifiers. The predictions of these base classifiers are subsequently integrated through an AdaBoost-based meta-classifier to generate the final classification decision.

Under the adopted stratified five-fold cross-validation protocol on the DARWIN dataset, the proposed framework achieved an accuracy of 98.28%, a precision of 97.78%, a recall of 98.88%, an F1-score of 98.32%, and an AUC of 0.94, indicating strong discriminative performance. In addition, evaluation on an independent handwriting-signature cohort achieved an accuracy of 91.76%, with a sensitivity of 90.74% and a specificity of 93.55%. These findings provide additional evidence that the proposed framework retains substantial discriminative capability beyond the dataset used for model development.

Nevertheless, the reported results should be interpreted with caution. The DARWIN dataset is relatively small, and the performance estimates remain dependent on the characteristics of the available datasets and the adopted experimental protocols. Moreover, although the independent evaluation provides additional evidence regarding generalizability, further validation on larger, more diverse, and clinically representative cohorts is required before conclusions regarding clinical applicability or deployment can be drawn.

The proposed framework should therefore be considered a promising approach for handwriting-based AD classification and potential screening rather than as a validated clinical diagnostic system or a demonstration of prospective early detection. Future work will focus on validation across additional independent cohorts, integration of multimodal biomarkers, and the application of explainable artificial intelligence techniques. Although explainability has already been investigated in handwriting-based AD classification, future research will explore how existing XAI approaches can be extended to interpret the learned representations and predictions of the proposed hybrid CNN-ensemble framework.

Author Contributions: Conceptualization, N.D.; methodology, I.M.; software, N.D. and S.T.; validation, N.D., S.T. and M.S.B.; formal analysis, S.T.; investigation, N.D.; resources, M.A. and A.H.; data curation, N.D.; writing—original draft preparation, N.D. and S.T.; visualization, N.D. and S.T.; supervision, M.A. and A.H.; project administration, M.A. and M.S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Dataset 1: <https://doi.org/10.24432/C55D0K> and Dataset 2: <https://doi.org/10.1007/s40520-023-02523-7>, all accessed on 1 September 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alzheimer's Association. 2025 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **2025**, *21*, e70235. [\[CrossRef\]](#)
2. Jack, C.R., Jr.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Dunn, B.; Haeberlein, S.B.; Holtzman, D.M.; Jagust, W.; Jessen, F.; Karlawish, J.; et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's Dement.* **2018**, *14*, 535–562. [\[CrossRef\]](#)
3. Titouni, S.; Djefal, N.; Belazzoug, M.; Hammache, B.; Messaoudene, I.; Hedir, A. A Confidence-Gated Hybrid CNN Ensemble for Accurate Detection of Parkinson's Disease Using Speech Analysis. *Electronics* **2026**, *15*, 587. [\[CrossRef\]](#)
4. Huang, G.; Li, R.; Bai, Q.; Alty, J. Multimodal learning of clinically accessible tests to aid diagnosis of neurodegenerative disorders: A scoping review. *Health Inf. Sci. Syst.* **2023**, *11*, 32. [\[CrossRef\]](#)
5. Domínguez-Fernández, C.; Egiguren-Ortiz, J.; Razquin, J.; Gómez-Galán, M.; Heras-García, L.D.L.; Paredes-Rodríguez, E.; Astigarraga, E.; Miguélez, C.; Barreda-Gómez, G. Review of technological challenges in personalised medicine and early diagnosis of neurodegenerative disorders. *Int. J. Mol. Sci.* **2023**, *24*, 3321. [\[CrossRef\]](#)
6. Tukur, H.N.; Uwishema, O.; Akbay, H.; Sheikah, D.; Correia, I.F.S. AI-assisted ophthalmic imaging for early detection of neurodegenerative diseases. *Int. J. Emerg. Med.* **2025**, *18*, 90. [\[CrossRef\]](#)
7. McKhann, G.M.; Knopman, D.S.; Chertkow, H.; Hyman, B.T.; Jack, C.R., Jr.; Kawas, C.H.; Klunk, W.E.; Koroshetz, W.J.; Manly, J.J.; Mayeux, R.; 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 Dement.* **2011**, *7*, 263–269. [\[CrossRef\]](#)
8. Hansson, O. Biomarkers for neurodegenerative diseases. *Nat. Med.* **2021**, *27*, 954–963. [\[CrossRef\]](#)
9. Impedovo, D.; Pirlo, G. Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective. *IEEE Rev. Biomed. Eng.* **2019**, *12*, 209–220. [\[CrossRef\]](#)
10. Vessio, G. Dynamic Handwriting Analysis for Neurodegenerative Disease Assessment: A Literary Review. *Appl. Sci.* **2019**, *9*, 4666. [\[CrossRef\]](#)
11. Kelley, B.J.; Petersen, R.C. Alzheimer's Disease and Mild Cognitive Impairment. *Neurol. Clin.* **2007**, *25*, 577–609. [\[CrossRef\]](#)
12. Rosenblum, S.; Werner, P. Assessing the handwriting process in healthy elderly persons using a computerized system. *Aging Clin. Exp. Res.* **2006**, *18*, 433–439. [\[CrossRef\]](#)
13. Yan, J.H.; Rountree, S.; Massman, P.; Doody, R.S.; Li, H. Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. *J. Psychiatr. Res.* **2008**, *42*, 1203–1212. [\[CrossRef\]](#)
14. Cilia, N.D.; De Gregorio, G.; De Stefano, C.; Fontanella, F.; Marcelli, A.; Parziale, A. Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Eng. Appl. Artif. Intell.* **2022**, *111*, 104822. [\[CrossRef\]](#)
15. De Gregorio, G.; Desiato, D.; Marcelli, A.; Polese, G. A Multi Classifier Approach for Supporting Alzheimer's Diagnosis Based on Handwriting Analysis. In *Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges*; Springer International Publishing: Cham, Switzerland, 2021; pp. 559–574. [\[CrossRef\]](#)
16. Mitra, U.; Rehman, S. ML-Powered Handwriting Analysis for Early Detection of Alzheimer's Disease. *IEEE Access* **2024**, *12*, 69031–69050. [\[CrossRef\]](#)
17. Ul Rehman, S.; Mitra, U. Cognitive Handwriting Insights for Alzheimer's Diagnosis: A Hybrid Framework. *Information* **2025**, *16*, 249. [\[CrossRef\]](#)
18. Parziale, A.; Della Cioppa, A.; Marcelli, A. Investigating One-Class Classifiers to Diagnose Alzheimer's Disease from Handwriting. In *Proceedings of the Image Analysis and Processing—ICIAP 2022*; Springer International Publishing: Cham, Switzerland, 2022; pp. 111–123. [\[CrossRef\]](#)
19. Rajendran, S.; R, N.; Selvadass, M. Hybrid Machine Learning Model Using Particle Swarm Optimization for Effectual Diagnosis of Alzheimer's Disease from Handwriting. In *Proceedings of the 2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, Bangalore, India, 21–23 December 2022; pp. 491–495. [\[CrossRef\]](#)
20. Demircioglu Diren, D. Design and Validation of a Hybrid Machine Learning Model for Alzheimer's Detection Using Handwriting Data. *Am. J. Alzheimer's Dis. Other Dement.* **2025**, *40*, 1–20. [\[CrossRef\]](#)
21. Öcal, H. A Novel Approach to Detection of Alzheimer's Disease from Handwriting: Triple Ensemble Learning Model. *Gazi Üniversitesi Fen Bilim. Derg. Part C Tasar. VE Teknol.* **2024**, *12*, 214–223. [\[CrossRef\]](#)
22. Erdogmus, P.; Kabakus, A.T. The promise of convolutional neural networks for the early diagnosis of the Alzheimer's disease. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106254. [\[CrossRef\]](#)

23. Gattulli, V.; Impedovo, D.; Pirlo, G.; Semeraro, G. Handwriting Task-Selection based on the Analysis of Patterns in Classification Results on Alzheimer Dataset. In Proceedings of the Workshop on Data Science Techniques for Datasets on Mental and Neurodegenerative Disorders (DSTNDS 2023), Zürich, Switzerland, 22 June 2023.
24. Önder, M.; Şentürk; Polat, K.; Paulraj, D. Diagnosis of Alzheimer's Disease Using Boosting Classification Algorithms. In Proceedings of the 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 1–2 November 2023; pp. 1–5. [\[CrossRef\]](#)
25. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [\[CrossRef\]](#)
26. Wilyanto, N.; Alamsyah, D. Handwriting-Based Alzheimer's Disease Detection Using VGG-19. *J. Appl. Inform. Comput.* **2026**, *10*, 1239–1252. [\[CrossRef\]](#)
27. Ho, N.T.N.; Gonzalez, P.; Gogovi, G.K. Writing the Signs: An Explainable Machine Learning Approach for Alzheimer's Disease Classification from Handwriting. *Healthc. Technol. Lett.* **2025**, *12*, e70006. [\[CrossRef\]](#)
28. Bone, M.B.; Freedman, M.; Black, S.E.; Felsky, D.; Kumar, S.; Pugh, B.; Strother, S.C.; Tang-Wai, D.F.; Tartaglia, M.C.; Buchsbaum, B.R. A Vision Transformer Approach for Fully Automated and Scalable Dementia Screening Using Clock Drawing Test Images. *Alzheimer's Dement. Diagn. Assess. Dis. Monit.* **2025**, *17*, e70171. [\[CrossRef\]](#)
29. Gong, C.; Qin, H.; El-Yacoubi, M.A. Hybrid Transformer for Early Alzheimer's Detection: Integration of Handwriting-Based 2D Images and 1D Signal Features. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 8293–8305. [\[CrossRef\]](#)
30. Sharma, R.; Acharya, V. Lightweight Vision Transformer with Transfer Learning for Interpretable Alzheimer's Disease Severity Assessment. *Sci. Rep.* **2025**, *15*, 44028. [\[CrossRef\]](#)
31. Li, L.; Ma, J.; Zou, B.; Liao, S.; Zhi, L. Explainable graph network with information fusion of data and medical knowledge for KOA diagnosis. *Biomed. Signal Process. Control* **2026**, *127*, 111097. [\[CrossRef\]](#)
32. Li, L.; Ma, J.; Zou, B.; Xu, H.; Liao, S.; Xiong, W.; Zhi, L. SymUnet-DynCFC: Multimodal MRI fusion for robust cartilage segmentation and clinically confirmed moderate-to-severe KOA diagnosis. *Inf. Fusion* **2026**, *130*, 104145. [\[CrossRef\]](#)
33. Li, L.; Xing, Y.; Xiong, W.; Liao, S.; Zou, B.; Sun, X.; Zhi, L. SwMrNet: A Multi-Target Tissue Segmentation Method for Robust and Accurate Clinical Knee Diagnosis Assistance. *Bioengineering* **2026**, *13*, 784. [\[CrossRef\]](#)
34. Pereira, C.A.; Wen, C.L.; Miguel, E.C.; Polanczyk, G.V. A Randomised Controlled Trial of a Web-Based Educational Program in Child Mental Health for Schoolteachers. *Eur. Child Adolesc. Psychiatry* **2015**, *24*, 931–940. [\[CrossRef\]](#)
35. Naomi, P.A.; Lorenzo, D.; Rita, C.; Francesca, P.; Teresa, D.; Giorgio, F.; Andrea, A.; Giuseppe, S.; Stefano, Z.; Nadia, B. Does cognitive decline influence signing? *Aging Clin. Exp. Res.* **2023**, *35*, 2685–2691. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.