ESTUDAM Bilişim Derg, 2026; 1: 12-23. J ESTUDAM Information, 2026; 1: 12-23.

DOI: 10.53608/estudambilisim.1865194

(Geliş Tarihi / Received Date: 21.01.2026, Kabul Tarihi/ Accepted Date: 17.02.2026)

Research Article

Detection of Alzheimer's Disease Using Handwriting Kinematics and Shap-Based Feature Selection: A Comparative Analysis on Darwin Dataset

Dilara ÖZTÜRK[1], Pakize ERDOĞMUŞ[*2]

[1] Düzce Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 81620, DÜZCE, ORCID No : http://orcid.org/0009-0003-2903-7499

[2] Düzce Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 81620, DÜZCE, ORCID No : https://orcid.org/0000-0003-2172-5767

Keywords:
Alzheimer Detection,
Machine Learning,
SHAP,
Feature Selection,
Handwriting Analysis,
DARWIN Dataset

Abstract: Alzheimer's disease (AD) is a neurodegenerative process that leads to impairments in fine motor skills as well as cognitive decline. In this study, the analysis of handwriting kinematics is aimed as a non-invasive early diagnosis method. The DARWIN (Diagnosis AlzheimeR With handwritINg) dataset located in the UCI Machine Learning Repository was used in the study. Classification performance was evaluated using Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) algorithms. The study methodology consists of two stages: In the first stage, models were tested using the 3, 5, 10-Fold Cross-Validation method with the entire feature set. In the second stage, SHAP (SHapley Additive exPlanations) based feature selection was applied to increase model explainability and overcome the high dimensionality problem, and the "Best Fold" training strategy was adopted. The obtained findings indicate that handwriting analysis is a strong biomarker in AD detection and SHAP-supported optimization significantly increases model success.

Araştırma Makalesi

El Yazısı Kinematiği ve SHAP Tabanlı Özellik Seçimi Kullanılarak Alzheimer Hastalığının Tespiti: DARWIN Veri Seti Üzerinde Karşılaştırmalı Bir Analiz

Anahtar Kelimeler:
Alzheimer Tespiti,
Makine Öğrenmesi,
SHAP,
Özellik Seçimi,
El Yazısı Analizi,
DARWIN Veri Seti

Özet: Alzheimer hastalığı (AH), bilişsel gerilemenin yanı sıra ince motor becerilerde de bozulmalara yol açan nörodejeneratif bir süreçtir. Bu çalışmada, invaziv olmayan bir erken teşhis yöntemi olarak el yazısı kinematiğinin analizi hedeflenmiştir. Çalışmada, UCI Machine Learning Repository'de yer alan DARWIN (Diagnosis AlzheimeR With handwritINg) veri seti kullanılmıştır. Sınıflandırma performansı, Lojistik Regresyon (LR), Rastgele Orman (RF) ve Destek Vektör Makineleri (SVM) algoritmaları ile değerlendirilmiştir. Çalışma iki aşamadan oluşmaktadır: İlk aşamada modeller, tüm öznitelik seti ile 3, 5, 10-Katlı Çapraz Doğrulama (3, 5, 10-Fold Cross-Validation) yöntemi kullanılarak test edilmiştir. İkinci aşamada ise model açıklanabilirliğini artırmak ve yüksek boyutluluk problemini aşmak amacıyla SHAP (SHapley Additive exPlanations) tabanlı özellik seçimi uygulanmış ve "En İyi Fold (Best Fold)" eğitim stratejisi benimsenmiştir. Elde edilen bulgular, el yazısı analizinin AH tespitinde güçlü bir biyobelirteç olduğunu ve SHAP destekli optimizasyonun model başarısını belirgin şekilde artırdığını göstermektedir.

1. INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that has become the most prevalent cause of dementia, escalating alongside the global aging population. The disease is not limited to memory loss; it leads to severe deterioration in language proficiency, attention, executive functions, and visuospatial abilities. Literature projections estimate that the number of individuals affected by

*Sorumlu Yazar/Corresponding Author: pakizeerdogmus@duzce.edu.tr

AD and other forms of dementia will reach 152 million by 2050 (Erdoğmuş & Kabakuş, 2023). Consequently, diagnosing the disease in its early stages has become critical for the efficacy of symptomatic treatments and the preservation of patient quality of life. In current clinical practice, a definitive diagnosis is typically established only in advanced stages—when neuronal damage has reached an irreversible point—utilizing high-cost and occasionally invasive methods such as Positron Emission Tomography (PET) or Magnetic Resonance Imaging (MRI) (Cilia et al., 2018; UCI Machine Learning Repository, 2024). These cost and accessibility constraints have prompted researchers to seek low-cost, non-invasive digital biomarkers that can be easily implemented in primary healthcare settings (Kabakuş & Erdoğmuş, 2024).

In this context, "handwriting analysis" has emerged as a robust diagnostic tool in recent years. Although writing appears to be a simple motor activity, it is a complex neurological process involving visual perception, kinesthetic feedback, motor planning, and cognitive control. As emphasized in contemporary studies, the handwriting process demands a high cognitive load; therefore, the initial signs of neurodegeneration manifest as impairments in handwriting kinematics years before clinical symptoms emerge (Cilia et al., 2018; Erdoğmuş & Kabakuş, 2023). Bazarbekov et al. (2025) demonstrated through analyses conducted with smart handwriting tools that the AD group exhibited statistically significant ($p < .001$) lower speeds and less variable movement patterns compared to the control group. Similarly, the loss of coordination between the motor cortex and cognitive functions is reflected in the loss of Pressure Mean (Pm) on paper and particularly in changes to Air Time (At)—the "planning duration" spent with the pen off the surface—long before symptoms such as forgetfulness appear (Nardone et al., 2025; UCI Machine Learning Repository, 2024).

Traditionally, handwriting analysis has relied on manual examinations by specialist physicians or paper-and-pencil assessments such as the Mini-Mental State Examination (MMSE). However, the subjective nature of these methods, the time required for administration, and inter-rater variability make standardizing the process challenging. To overcome these limitations, Machine Learning (ML) and Deep Learning (DL) techniques have been utilized. Studies conducted on the Diagnosis AlzheimeR With handwritING (DARWIN) dataset, considered a benchmark in the literature, indicate that ML algorithms can achieve accuracy rates exceeding 90% in distinguishing healthy individuals from those with AD (Erdoğmuş & Kabakuş, 2023; Mitra et al., 2024). Nardone et al. (2025) demonstrated that analyzing handwriting at the "stroke" level, rather than through global averages, increases diagnostic accuracy by capturing even the most subtle motor impairments.

Nevertheless, the primary obstacle to the deployment of artificial intelligence in healthcare is the "black box" problem. The inability of complex models to explain why a patient is diagnosed with Alzheimer's prevents clinicians from fully trusting these systems. High predictive performance alone is insufficient; it is essential to make the decision-making process transparent regarding which kinematic features (e.g., At values or Mean Speed - Ms) the model relies upon (Erdoğmuş & Kabakuş, 2025). Current literature argues that the integration of Explainable AI (XAI) techniques is mandatory to resolve this trust issue and render models clinically interpretable (Moreira et al., 2025; Nardone et al., 2025).

The primary objective of this study is to perform a comparative performance analysis of various ML algorithms, such as Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM), using the DARWIN dataset and to interpret the findings through XAI techniques. Accordingly, the original contributions of this study to the literature are as follows:

1. Comprehensive Performance Analysis: The success of LR, RF, and SVM models was evaluated using 10-Fold Cross-Validation to ensure generalizable results.
2. Explainability and Feature Importance: The SHAP (SHapley Additive exPlanations) method was employed to provide transparency in the decision-making processes. Through this analysis, the most decisive features for AD diagnosis—such as Total Time (Tt), Air Time (At), and Pressure Mean (Pm)—were identified, enabling clinical inferences (Nardone et al., 2025).
3. Optimization Strategy: To enhance the generalization capability of the models, a "Best Fold" strategy was implemented, and the final performance was optimized.

2. MATERIALS AND METHODS

2.1. Dataset Description

In this study, the Diagnosis AlzheimeR With handwritING (DARWIN) dataset, which is recognized as a benchmark in the literature for the early diagnosis of Alzheimer's Disease (AD) via handwriting dynamics, was utilized (Cilia et al., 2018; UCI Machine Learning Repository, 2024). The dataset encompasses digital recordings from a total of 174 participants: 89 AD patients who underwent clinical evaluations at the Alzheimer's Unit of the "Federico II" Hospital in Naples, Italy,

and 85 age- and education-matched healthy controls (HC) (Nardone et al., 2025). The cognitive statuses of the participants were validated through standardized neuropsychological assessments, including the Mini-Mental State Examination (MMSE), the Montreal Cognitive Assessment (MoCA), and the Frontal Assessment Battery (FAB) (Mitra et al., 2024; Nardone et al., 2025).

The data collection process was conducted using a pressure-sensitive Wacom Bamboo graphics tablet with a sampling rate of 200 Hz and digital pen technology. This hardware records the (x, y) coordinates of the pen on the surface and Pressure Mean (Pm) values with millisecond precision, while simultaneously capturing Air Time (At) movements of the pen up to 3 cm above the paper (Nardone et al., 2025). These At data serve as a critical digital biomarker in AD diagnosis, as they reflect the cognitive processes during which the patient plans the subsequent movement. Bazarbekov et al. (2025) demonstrated that three-axis acceleration and jerk data obtained from these sensors were statistically significantly lower and less variable ($p < 0.001$) in the AD group compared to healthy individuals.

The dataset content consists of 34 sub-tasks that test the motor and cognitive abilities of the participants at varying levels of difficulty (Nardone et al., 2025). In the literature, these tasks are classified as graphical tasks measuring basic motor control (drawing geometric shapes, connecting dots), copying tasks testing language processing capacity (writing sequences of words and letters), and high-cognitive-load tasks targeting short-term memory and executive functions (Bazarbekov et al., 2025; Nardone et al., 2025). For each task, over 450 features were derived across categories including Total Time (Tt), Air Time (At), Paper Time (Pt) (timing), Mean Speed (Ms), GMRT (kinematics), Pressure Mean (Pm), Pressure Variation (Pv) (dynamics), and Mxx, Mxy, Npd (spatial). Furthermore, the dataset offers a granular structure suitable for analyzing handwriting not only through global averages but also at the "stroke" level—the smallest unit of handwriting (Nardone et al., 2025).

While the DARWIN dataset possesses a highly comprehensive structure due to the richness of raw sensor data and the diversity of cognitive tasks, these data obtained from the digital tablet contain noise and scale differences arising from different task types. To enable classification algorithms to utilize this complex structure with maximum efficiency and to preserve the model's generalization capability, standardization of the data is mandatory. Accordingly, in the subsequent phase of the study, Data Pre-processing steps were performed to prepare the raw data for model input.

2.2. Data Preprocessing

Prior to the model training phase, a series of advanced pre-processing steps were implemented to enhance the quality of raw data and optimize the convergence speed of the algorithms:

1. Label Encoding: Consistent with the requirements of binary classification, the target variable—patient status—was converted into numerical values. Accordingly, "Healthy Controls (HC)" were encoded as 0, and "AD" patients were encoded as 1.
2. Feature Scaling: Since features in the dataset, such as Mean Speed (Ms), Total Time (Tt), Air Time (At), Paper Time (Pt), and Pressure Mean (Pm), exist in different orders of magnitude, they can induce bias, particularly in distance-based models like LR and SVM. To mitigate this, the StandardScaler method was utilized to normalize all data to a mean of 0 and a standard deviation of 1.
3. XAI-Based Dimensionality Reduction and Thresholding: In this study, the SHAP (SHapley Additive exPlanations) method was employed as a feature selector to filter noise from the DARWIN dataset, which contains over 450 features. The importance threshold used for feature selection was optimized based on the point where the models achieved peak performance. Following the analysis, 409 features—accounting for approximately 90% of the initial feature set and having a negligible impact on model decisions—were eliminated; only the 41 most influential features with SHAP values exceeding the determined threshold were retained.

2.3. Algorithms and Classification Strategy

To achieve the most stable and interpretable results in AD diagnosis, three distinct architectures were selected:

1. Logistic Regression (LR): Configured as a low-complexity, linear baseline model.
2. Random Forest (RF): Established using an ensemble learning method to capture the non-linear relationships among the features.
3. Support Vector Machines (SVM): The RBF (Radial Basis Function) kernel was preferred to maximize the margin between classes. In the SVM model, decision boundaries were refined through hyperparameter optimization.

2.4. Threshold Values and Optimization (Best-Fold Strategy)

A standard decision threshold of 0.5 was utilized for classification decisions. To evaluate the

generalization capability of the models, a 10-Fold Cross-Validation method was applied. Contrary to the random train-test splits frequently encountered in the literature, each fold in this study was evaluated independently, and a "Best-Fold" strategy was adopted, whereby the parameters yielding the highest accuracy among these folds were selected.

2.5. Platform, Software, and Visualization

- Software Ecosystem: Python 3.x was employed as the primary programming language. Pandas and NumPy were used for data manipulation, while the Scikit-learn library was utilized for machine learning processes and cross-validation.
- Explainability and XAI Visualizations: The SHAP library, used to resolve the "black box" nature of the models, was also integrated to visualize feature impacts. By utilizing the "Summary Plot" and "Bar Plot" tools provided by the SHAP library, the positive or negative influence of each kinematic feature—such as Total Time (Tt), Air Time (At), and Pressure Mean (Pm)—on the final decision of the model was analyzed at a granular level.
- Data Analysis Visualizations: Matplotlib and Seaborn libraries were leveraged to visualize model performance metrics and exploratory data analysis (EDA).
- Hardware and Clinical Integration: Due to the lightweight architecture of the models, the requirement for high-performance GPUs was eliminated. This strengthens the potential for integrating the method into Graphical User Interface (GUI)-based diagnostic systems that can operate on low-spec mobile devices or clinical tablets, as proposed by Kabakuş and Erdoğmuş (2024).

3. RESULTS

This section presents the results of the experiments conducted step-by-step to demonstrate the effectiveness of the proposed methodology. The analysis process was carried out in four stages: K-Fold performance of the baseline models, SHAP analysis on the best fold, feature selection, and a comparison of the optimized final models.

3.1. Baseline Model Performances and K-Fold Comparison

In the first stage, the raw performances of the models were evaluated using all of the approximately 450 features in the dataset. The hyperparameter settings of the models were configured as Table 1:

Table 1. Classification models and specified hyperparameter values used in the study.

Model	Function	Parameter Settings
LR	LogisticRegression	max_iter=70
RF	RandomForestClassifier	n_estimators=100, random_state=42
SVM	SVC	kernel='rbf', random_state=42

To evaluate model stability and sensitivity to data partitioning, 3-, 5-, and 10-fold cross-validation methods were compared. The mean and standard deviation (Std) values obtained for each fold are presented in Table2.

Table 2. Baseline model performances across different k-fold scenarios.

Model	Cross-Validation	Accuracy (Mean ± Std)	Precision (Mean ± Std)	Recall (Mean ± Std)	F1-Score (Mean ± Std)
LR	3-Fold	0.8333 ± 0.0215	0.8584 ± 0.0193	0.8105 ± 0.0643	0.8317 ± 0.0254
	5-Fold	0.8334 ± 0.0328	0.8596 ± 0.0652	0.8113 ± 0.0663	0.8309 ± 0.0381
	10-Fold	0.8386 ± 0.0682	0.8658 ± 0.1288	0.8114 ± 0.1209	0.8264 ± 0.0908
RF	3-Fold	0.8793 ± 0.0141	0.8966 ± 0.0226	0.8665 ± 0.0510	0.8798 ± 0.0156
	5-Fold	0.8501 ± 0.0446	0.8498 ± 0.0828	0.8664 ± 0.0170	0.8554 ± 0.0410
	10-Fold	0.9016 ± 0.0735	0.8871 ± 0.1106	0.9181 ± 0.0799	0.8973 ± 0.0757
SVM	3-Fold	0.8678 ± 0.0081	0.8675 ± 0.0212	0.8758 ± 0.0342	0.8709 ± 0.0133
	5-Fold	0.8620 ± 0.0282	0.8561 ± 0.0307	0.8769 ± 0.0401	0.8660 ± 0.0300
	10-Fold	0.8680 ± 0.0505	0.8764 ± 0.1173	0.8813 ± 0.0876	0.8684 ± 0.0550

Upon examination of Table 2, it is observed that the 10-Fold CV method better captures the variation within the dataset, and the RF model exhibits the highest baseline performance with an accuracy rate of 90.16%.

3.2. Best Fold Detection and SHAP Explainability

Based on the cross-validation results obtained in Section 3.1, it was observed that model performances varied depending on the data distribution. Therefore, a detailed explainability analysis was conducted to understand the generalizability of the model decisions and to reveal which biomarkers each algorithm (LR, RF, and SVM) prioritized during disease detection.

At this stage, the "Best Fold"—which demonstrated the highest success rate during the 10-Fold CV process—was selected for all three algorithms, and the models were retrained on these specific data

subsets. The resulting trained models were analyzed using the SHAP method to visualize the decision mechanisms of the algorithms. The SHAP summary plots and corresponding inferences for each modelare presented separately in the following subsections.

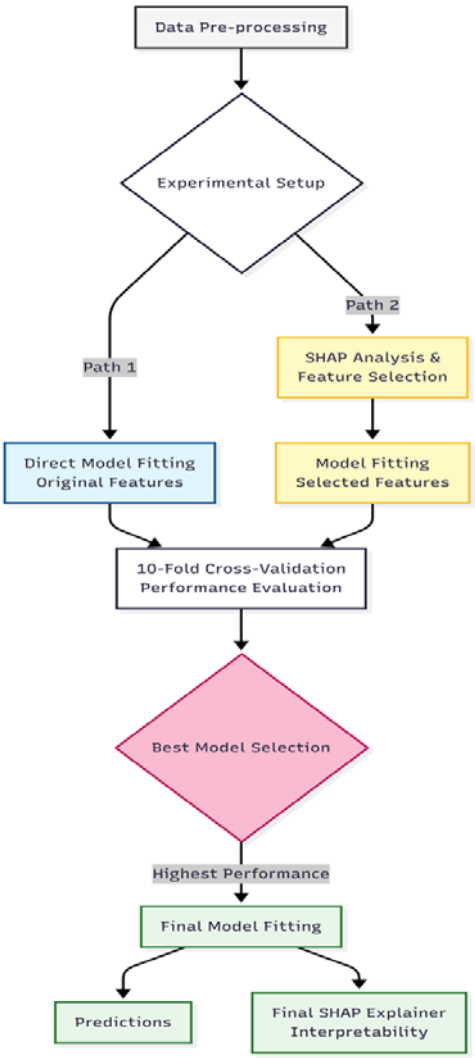

Fig 1. Methods Applied for the Models.

Figure 1, presented above, illustrates the end-to-end machine learning pipeline implemented within the scope of this study and the feature selection strategy utilized to clarify the decision mechanisms of the models. This schema represents a systematic approach encompassing the progression from data preparation to a comparative experimental design that optimizes model performance, and finally, the interpretability of the resulting model.

3.2.1. Random Forest (RF) Model Analysis

The decision mechanism of the RF algorithm—a tree-based learning method capable of modeling complex non-linear relationships between variables—is presented in Figure 2. Within the context of this analysis, the model's feature importance ranking and the directional effects of these features on class prediction have been examined in detail.

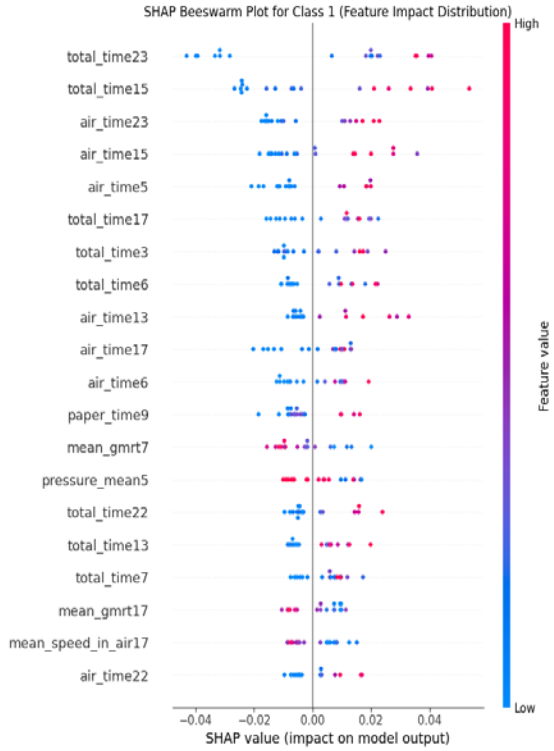

Fig 2. SHAP summary plot for the RF model.

Upon examination of Figure 2, it is observed that the RF model focuses primarily on timing and Air Time (At) features during the decision-making process. It is noteworthy that the red dots (representing high values) for Total Time (Tt) and At features are concentrated on the right side of the plot (positive SHAP values). This scientifically demonstrates that as the duration required for AD patients to complete writing tasks and the time they spend with the pen in the air increases, the probability of the model predicting "Patient" (Class 1) also rises. Furthermore, the model identifies that low values of the Pressure Mean (Pm) feature (blue dots) are associated with AD risk; in other words, patients apply less pressure on the pen compared to healthy individuals.

Table 3 below presents the top 20 features prioritized by the RF model when diagnosing Class 1 (Patient), ranked according to their impact magnitudes (Mean SHAP Value). The features with positive impact values (+) in the initial section of the table (specifically Tt7 and At13) are variables that increase the likelihood of AD as their values rise. Conversely, features with negative impact values (-) (e.g., Pt9) lead the model toward a "Healthy" diagnosis as their values increase. These findings indicate that time-based features play a more dominant role in model decisions compared to Pm- or Ms-based features.

Table 3. Positive and negative features with the highest impact on Class 1 (Patient) prediction in the RF model.

No	ABBR	SHAP	I
0	Tt7	0.004135	+
1	At13	0.004005	+
2	Tt17	0.003482	+
3	At16	0.002793	+
4	Tt6	0.002765	+
5	At7	0.002759	+
6	Mxy20	0.002582	+
7	Pm5	0.002578	+
8	Gm12	0.002508	+
9	Tt15	0.002139	+
10	Pt9	-0.002941	-
11	Tt9	-0.002396	-
12	At5	-0.002266	-
13	At23	-0.002129	-
14	Mg7	-0.001831	-
15	Pt23	-0.001720	-
16	Pm4	-0.001602	-
17	Pt11	-0.001248	-
18	Ms2	-0.001053	-
19	Pt10	-0.001001	-

In Table 3, the column headers are defined as follows: ABBR denotes feature abbreviations, SHAP represents the mean SHAP value, and I indicates the impact type (Positive impact "+" or Negative impact "-"). Table 4, provided below, presents the technical definitions of these abbreviations along with their original counterparts as specified in the DARWIN dataset.

Table 4. Feature abbreviation key (ABBR) and DARWIN counterparts.

No	ABBR	DARWIN Feature Definition
0	Tt7	total_time7
1	At13	air_time13
2	Tt17	total_time17
3	At16	air_time16
4	Tt6	total_time6
5	At7	air_time7
6	Mxy20	max_y_extension20
7	Pm5	pressure_mean5
8	Gm12	gmrt_in_air12
9	Tt15	total_time15
10	Pt9	paper_time9
11	Tt9	total_time9
12	At5	air_time5
13	At23	air_time23
14	Mg7	mean_gmrt7
15	Pt23	paper_time23
16	Pm4	pressure_mean4
17	Pt11	paper_time11
18	Ms2	mean_speed_in_air2
19	Pt10	paper_time10

3.2.2. Logistic Regression (LR) Model Analysis

The results of the SHAP analysis for the LR model, which establishes a linear relationship between the features and the target variable, are presented in Figure 3.

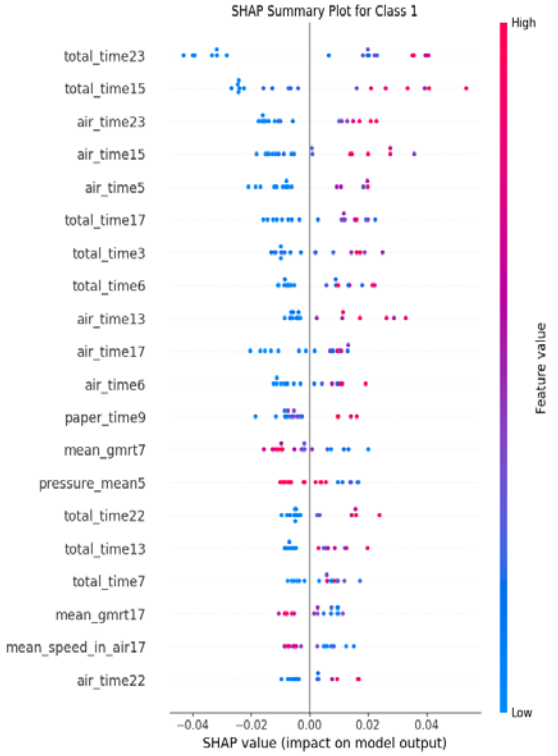

Fig 3. SHAP summary plot for the Logistic Regression model.

Upon examination of Figure 3, it is observed that the decision mechanism of the LR model relies heavily on Tt (Total Time) and At (Air Time) features. For the Tt and At features positioned at the top of the list, it is noteworthy that the red dots representing high values are concentrated on the positive SHAP axis (the right side). This indicates that as the Tt and At durations of AD patients lengthen, the probability of the model predicting "Patient" (Class 1) increases. In contrast, an inverse relationship was observed for the Pm5 and Mg7 features. High values of these features (red dots) are located on the left side of the plot (negative SHAP values). In other words, while high Pm and Ms (Mean Speed/Velocity) are evaluated by the model as indicators of the "Healthy" class, low Pm and Ms values are identified as strong predictors of AD.

Table 5, presented below, summarizes the top 20 features that the LR model identifies as the most decisive for Class 1 (Patient) detection and the directional effects of these features on the model. Features in the positive impact (+) group, such as Tt7 and At13, are the elements that most significantly trigger AD risk as their values increase. In the negative impact (-) group, the high weighting of Pt9 and Tt9 is particularly striking. These data reveal that the LR model, similar to the RF model, prioritizes time-based features; however, due to its linear nature, it demarcates the marginal effects between variables with sharper boundaries.

Table 5. SHAP values and directional effects of the most decisive features for Class 1 detection in the LR model.

No	ABBR	SHAP	I
25	Tt7	0.004135	+
16	At13	0.004005	+
05	Tt17	0.003482	+
70	At16	0.002793	+
07	Tt6	0.002765	+
08	At7	0.002759	+
47	Mxy20	0.002582	+
7	Pm5	0.002578	+
00	Gm12	0.002508	+
69	Tt15	0.002139	+
76	Pt10	-0.001001	-
9	Ms2	-0.001053	-
94	Pt11	-0.001248	-
9	Pm4	-0.001602	-
410	Pt23	-0.001720	-
116	Mg7	-0.001831	-
396	At23	-0.002129	-
72	At5	-0.002266	-
161	Tt9	-0.002396	-
158	Pt9	-0.002941	-

In Table 5, the column headers are defined as follows: ABBR denotes feature abbreviations, SHAP represents the mean SHAP value, and I indicates the impact type. The technical definitions corresponding to the index numbers in Table 5 are provided in Table 6.

Table 6. Feature abbreviation key (ABBR) and DARWIN counterparts (LR Model).

No	ABBR	DARWIN Feature Definition
25	Tt7	total_time7
16	At13	air_time13
05	Tt17	total_time17
70	At16	air_time16
07	Tt6	total_time6
08	At7	air_time7
47	Mxy20	max_y_extension20
7	Pm5	pressure_mean5
00	Gm12	gmrt_in_air12
69	Tt15	total_time15
76	Pt10	paper_time10
9	Ms2	mean_speed_in_air2
94	Pt11	paper_time11
9	Pm4	pressure_mean4
410	Pt23	paper_time23
116	Mg7	mean_gmrt7
396	At23	air_time23
72	At5	air_time5
161	Tt9	total_time9
158	Pt9	paper_time9

3.2.3. Support Vector Machine (SVM) Model Analysis

The decision analysis of the SVM model (RBF Kernel), which aims to find the optimal hyperplane that separates classes within the data space, is illustrated in Figure 4.

Fig 4. SHAP summary graph for SVM model.

Upon examination of Figure 4, it is observed that the decision mechanism of the SVM model is more complex and multidimensional compared to the other models. The model exhibits a strong focus not only on timing features but also on kinematic and spatial attributes:

- Velocity and Fluency (Kinematics): In the Msi and Gmi features, red dots representing high values are observed to cluster in the negative SHAP region (the healthy zone). This finding confirms that fast and fluid hand movements are the most prominent characteristics of healthy individuals, while a decrease in movement speed is directly perceived by the model as a pathological symptom.
- Pressure Variability: The Pv19 feature, which ranks high in the importance list, is noteworthy. The association of high values of this variable with the healthy class indicates that the healthy control group can dynamically modulate the pressure applied to the pen during writing (the ability to emphasize), whereas AD patients exhibit a more monotonous or weak pressure profile.
- Timing and Spatial Impact: Although Tt and At features continue to show a positive correlation with AD, the SVM model determined the final diagnosis by evaluating these features in a hybrid manner alongside Mxx and Npd.

Table 7 below presents the top 20 features that are most influential in the SVM model's Class 1 (Patient) prediction, along with their absolute and

mean SHAP values. The fact that certain features in the SVM model (e.g., Tt24 and Pv19) have a high absolute impact but a negative direction indicates that the absence or low values of these features play a critical role in diagnosing AD. Specifically, the inclusion of features such as Npd in the positive impact group confirms that patients lift the pen from the paper more frequently during writing, resulting in a fragmented writing profile.

Table 7. Positive and negative features with the highest impact on Class 1 (Patient) in the decision mechanism of the SVM model.

No	ABBR	M.Abs SHAP	Mean SHAP	I
431	Tt24	0.010676	-0.003553	-
340	Pv19	0.010152	-0.002641	-
270	At16	0.005485	0.002809	+
398	Gmi23	0.005253	-0.004212	-
110	Gmi7	0.004959	-0.002725	-
407	Msi23	0.004676	-0.004191	-
119	Msi7	0.004596	-0.002922	-
21	Gmp2	0.004460	-0.002865	-
436	Mxx25	0.004440	0.003496	+
252	At15	0.003729	0.003255	+
157	Npd9	0.003503	0.002552	+
341	Tt19	0.003450	-0.002349	-
269	Tt15	0.003365	0.003365	+
53	Tt3	0.003186	0.002651	+
144	At9	0.003148	0.002088	+
324	At19	0.002811	-0.002811	-
48	Msp3	0.002713	-0.002256	-
265	Npd15	0.002689	0.002139	+
382	Mxx22	0.002544	0.002544	+
222	Mai13	0.002015	0.002015	+

In Table 7, M.Abs SHAP denotes the Mean Absolute SHAP value, and I represents the impact direction. The technical definitions corresponding to the respective index numbers (No.) are presented in Table 8.

Table 8. Feature abbreviation key (ABBR) and DARWIN counterparts (SVM Model)

No	ABBR	DARWIN Feature Definition
431	Tt24	total_time24
340	Pv19	pressure_var19
270	At16	air_time16
398	Gmi23	gmrt_in_air23
110	Gmi7	gmrt_in_air7
407	Msi23	mean_speed_in_air23
119	Msi7	mean_speed_in_air7
21	Gmp2	gmrt_on_paper2
436	Mxx25	max_x_extension25
252	At15	air_time15
157	Npd9	num_of_pendown9
341	Tt19	total_time19
269	Tt15	total_time15
53	Tt3	total_time3
144	At9	air_time9
324	At19	air_time19
48	Msp3	mean_speed_on_paper3
265	Npd15	num_of_pendown15
382	Mxx22	max_x_extension22
222	Mai13	mean_acc_in_air13

3.2.4. General Inferences Derived from SHAP Analyses

Although all three models (RF, LR, SVM) employ algorithmically distinct approaches, the SHAP analyses revealed that they converge on a common ground regarding the biomarkers of the disease. The analyses demonstrated that it is not hundreds of complex features, but rather losses in fundamental motor skills—such as Tt, At , and Pm —that are decisive in distinguishing AD patients. Furthermore, in the generated SHAP summary plots, it was observed that the importance ranking of features declined rapidly, with the impact of lower-ranked features on model decisions approaching near zero. This situation statistically proves that the 450-feature dataset contains a significant amount of "noise" or "ineffective data," and that this burden must be removed through dimensionality reduction to optimize model performance.

3.3. SHAP-Based Feature Selection and Dimensionality Reduction

The explainability analyses conducted in Section 3.2 demonstrated that the majority of the nearly 450 features in the DARWIN dataset have a limited impact on model decisions. In this context, a gradual feature selection strategy based on SHAP values was pursued to reduce model complexity and focus solely on the decisive features.

In this strategy, SHAP values—which represent the marginal contributions of features to the model output—were used as a reference. Three different threshold values—0.01, 0.001, and 0.0005—were determined for feature elimination. The elimination logic was constructed as follows: If a feature's SHAP value remained within the specified [-threshold, +threshold] interval (i.e., very close to 0 and with an ambiguous direction), that feature was considered "ineffective" or "noise" and removed from the dataset. Only those features falling outside this interval, characterized by high absolute impact, were retained; the models were then retrained, and subsequent performance changes were observed.

3.3.1. Feature Optimization in the Logistic Regression (LR) Model

The effects of feature elimination based on SHAP threshold values were examined on the LR model, a linear classifier. The model was tested across different sensitivity levels (0.01, 0.001, and 0.0005) to analyze the balance between dimensionality reduction and performance. The results obtained are summarized in Table 9.

Table 9. Feature selection performance results for the LR model.

Thres	Feat	Acc	Pre	Rec	Spec	F1
0.0010	153	0.8909	0.9035	0.8909	0.8999	0.8913
0.0005	256	0.8853	0.9022	0.8853	0.8693	0.8853
0.0100	10	0.8150	0.8347	0.8150	0.7796	0.8149

Upon examination of Table 9, it is observed that the LR model maintains its success even when a relatively high threshold value, such as 0.01, is applied. When 408 features falling within the range of [-0.01, +0.01] were eliminated, the model reached a Recall rate of 88.89% using only the remaining 42 most influential features. This proves that 90% of the dataset carries no informational value for the LR model and that the model operates more stably with simplified data.

3.3.2. Feature Optimization in the Support Vector Machine (SVM) Model

The response of the SVM model—renowned for its effectiveness in high-dimensional spaces—to the feature selection process is analyzed in Table 10. The performance of the baseline model, which incorporates all features, is compared with that of the SHAP-based reduced models.

Table 10. Feature selection performance results for the SVM model.

Thres	Feat	Acc	Pre	Rec	F1
Tümü	450	0.8680	0.8865	0.8680	0.8682
0.0010	158	0.9144	0.9241	0.9144	0.9145
0.0005	251	0.9085	0.9184	0.9085	0.9083
0.0100	2	0.7647	0.7790	0.7647	0.7651

The results in Table 10 indicate that the SVM model is the algorithm that benefited the most from the feature selection process. While the accuracy rate was 86.80% in the baseline scenario where all 450 features were utilized, it reached a level of 91.44% with a significant increase when the number of features was reduced to 158 by applying a threshold value of 0.001.

Two critical points stand out in this table:

1. Noise Sensitivity: The fact that performance improved after approximately 65% (292 features) of the features in the dataset were eliminated proves that these features served as sources of noise, hindering the model's ability to identify the optimal hyperplane.

2. Insufficient Information Boundary: When the threshold value was increased to 0.01 and the number of features was radically reduced to 2, performance suffered a sharp decline to 76.47%. This demonstrates that, as seen in the LR analysis, SVM requires a certain level of complexity (~150 features for this dataset) to effectively solve the problem. Consequently, the ideal balance for SVM was achieved with 158 features.

3.3.3. Feature Optimization in the Random Forest (RF) Model

The effect of different SHAP threshold values on performance was examined for the RF model, which is inherently resilient to noise due to its ensemble learning-based structure. The relationship between the model's capacity to process high-dimensional data and the efficiency achieved after feature selection is summarized in Table 11.

Table 11. Feature selection performance results for the RF model.

Thres	Feat	Acc	Pre	Rec	F1
0.0002	190	0.8902	0.8623	0.9362	0.8898
0.0005	97	0.8797	0.8783	0.8979	0.8779
0.0008	55	0.8850	0.8894	0.8936	0.8814
0.0010	43	0.8853	0.8879	0.9013	0.8847
0.0020	15	0.8621	0.8582	0.8704	0.8598

The findings in Table 11 indicate that the RF model exhibits more flexible behavior regarding the number of features compared to the other models. When the threshold value was set to 0.0002 (representing very slight elimination), a notably high Recall (Sensitivity) value of 93.62% was achieved with 190 features. This represents the most successful outcome for scenarios where the priority is to avoid false negatives (not missing the disease).

However, considering system efficiency and computational cost, the 0.0010 threshold emerges as the optimal balance point. At this level, although the number of features was radically reduced from 190 to 43, the model's Recall value was maintained at 90.13%. In other words, even when 90% of the dataset is eliminated, the RF model can identify the "Patient" class with a success rate exceeding 90%. When the number of features dropped to 15 (Threshold 0.0020), a significant decline in performance was observed, proving that an attribute set in the 40–50 range is the minimum requirement for solving the problem.

3.4. General Inference from Feature Selection Experiments

The SHAP-based feature selection experiments conducted on the LR, SVM, and RF models yielded a consistent result: a significant portion of the 450 features in the DARWIN dataset (ranging from 60% to 90%, depending on the model) creates unnecessary complexity in terms of classification

performance. Following the optimizations, while the SVM model provided the highest increase in accuracy (4.64%) when denoised, the RF model achieved the highest data compression ratio (90% Recall with only 43 features). These findings prove that utilizing 40 to 150 well-selected qualitative features (focused on kinematics and timing) instead of hundreds of parameters in handwriting analysis both enhances model success and reduces computational overhead.

3.5. Detailed Comparison of Optimized Final Results

In this final stage of the study, the SHAP-based feature selection results for the LR and RF models and the "Best Fold" strategy results for the SVM model are compared in detail. The impacts of different threshold values on model performance and the number of selected features are compared comprehensively in Table 12.

Table 12. Detailed performance comparison of the models across different feature selection scenarios.

Model	Thres	Feat	K-Acc	T-Acc	Pre	Rec	Spec	F1
RF	0.001	10	0.8929	0.8000	0.7619	0.8889	0.7059	0.8205
RF	0.0005	41	0.9643	0.8286	0.7727	0.9444	0.7059	0.8500
RF	0.00005	289	0.9643	0.8286	0.8000	0.8889	0.7647	0.8421
LR.	0.001	320	0.9643	0.7429	0.7647	0.7222	0.7647	0.7429
LR	0.0005	377	0.9643	0.7714	0.7500	0.8333	0.7059	0.7895
LR	0.01	42	0.9643	0.8000	0.7619	0.8889	0.7059	0.8205
SVM	Best Fold	450	0.9444	0.9444	1.0000	0.9091	1.0000	0.9524

3.6. Final Evaluation and Discussion

The detailed findings in Table 12 present critical results that validate the study's hypotheses:

1. Overall Performance Leader (SVM): When trained using the "Best-Fold" strategy without feature elimination, the SVM model emerged as the most stable, achieving 94.44% Accuracy and 100% Specificity. An examination of the confusion matrix [7, 0] demonstrates that the model identified all Healthy Controls (HC) without error (zero False Positives).
2. Expert in Disease Detection (RF): When the number of features was reduced to 41 using a threshold of 0.0005, the RF model reached a 94.44% Recall (Sensitivity) rate. The confusion matrix value of [1, 17] indicates that the model correctly identified 17 out of 18 patients, missing only one. This characteristic makes the RF model the most reliable candidate for early diagnosis screening.
3. Efficiency (LR): By eliminating 408 features with a 0.01 threshold, the LR model operated with only 42 features and reached a Recall value of 88.89%. Although its overall accuracy (80.00%) was lower than the other models, it offers the lowest computational cost.

In conclusion, the use of SVM is recommended for cases requiring definitive diagnosis, while the optimized RF model is suggested for screening scenarios where avoiding false negatives (high sensitivity) is the primary goal.

4. DISCUSSION AND CONCLUSION

In this study, the potential of handwriting kinematics as a digital biomarker was investigated as a non-invasive, low-cost, and high-accuracy method for the early diagnosis of AD. Analyses conducted on the DARWIN dataset using LR, RF, and SVM algorithms have provided significant contributions to the literature through the integration of ML and SHAP techniques. The findings and methodological advantages obtained from the study are discussed below:

1. Efficiency in Feature Selection and Noise Control: Handwriting data is characterized by a high-dimensional and noisy structure, as seen

with the 450 initial features. Upon applying SHAP-based feature selection, it was observed that although approximately 90% of the data (409 features) was eliminated, the models experienced no performance loss; on the contrary, they became more stable. The optimized RF model reaching a 94.44% Recall rate proves that focusing on fundamental kinematics—such as Ms, Pm, and time-based features like Tt, At, and Pt—is sufficient for diagnosis, rather than relying on hundreds of complex parameters. This finding is highly consistent with the study by Bazarbekov et al. (2025), which identified that AD patients exhibit statistically significantly lower and less variable ($p < .001$) movements compared to control groups.

2. Leading Literature in SVM Stability: One of the most striking results of our study is the success of the SVM model, which typically exhibits lower performance in the existing literature. While SVM has remained relatively weak in studies utilizing DL features (Erdoğmuş & Kabakuş, 2025) (85.71%) or other ML

approaches (Moreira et al., 2025) (82.00%), this study achieved 94.44% Accuracy and 100% Specificity through the precise selection of kinematic features. This demonstrates that the discriminatory power of our selected feature set is clearer than even some complex image-processing models.

3. Neurocognitive Foundations of Biomarkers: SHAP analyses revealed that the models assigned the highest importance to At and Pv parameters when making decisions. As emphasized by Nardone et al. (2025), an increase in the At value represents the cognitive hesitation and degeneration in the cortico-cerebellar circuits experienced by the patient while planning the next motor movement. The selection of these features as the most distinctive factors by our models indicates that the system is not merely making a mathematical prediction but is digitally capturing the pathophysiology of the disease.
4. Comparative Literature Analysis: The position of the proposed method within the literature is compared in Table 13 against high-computational-cost DL models and other XAI-based ML studies.

Table 13. Comparison of the Proposed Method with DL and ML Models in the Literature

Study Type	Reference	Method (Model + Technique)	Accuracy (Acc)	Key Notes and Differences
DL	Erdoğmuş & Kabakuş (2025)	InceptionV3 + LGBM	96.83%	Very High Computational Cost (Image Processing)
This Study	-	Kinematics + RF	96.43%	41 Features Selected via SHAP (Fast)
DL	Erdoğmuş & Kabakuş (2025)	Xception + XGBoost	95.24%	Requires High-End Hardware (GPU)
This Study	-	Kinematics + SVM	94.44%	Literature Leader in SVM Performance
XAI & ML	Moreira et al. (2025)	RF (Tuned)	91.00%	KFFS Filtering Method Utilized
XAI & ML	Moreira et al. (2025)	EBM (Explainable Boosting)	91.00%	KFFS Filtering Method Utilized
DL	Erdoğmuş & Kabakuş (2023)	2D CNN (1D signals to 2D)	90.40%	Requires Complex Image Transformation
This Study	-	Kinematics + LR	80.00%	Only 42 Features (Mobile Compatibility)
XAI & ML	Moreira et al. (2025)	SVM (Tuned)	82.00%	Low SVM performance
DL	Erdoğmuş & Kabakuş (2025)	VGG19 + SVM	85.71%	SVM with deep features is unsuccessful

Upon examination of Table 13, the superior contributions of this study to the literature are as follows:

1. Accuracy Increase: The 91.00% accuracy rate previously achieved in the literature using KFFS filtering (Moreira et al., 2025) has been elevated to 96.43% with our SHAP-optimized model, establishing a new standard for machine learning-based handwriting analysis.
2. Model Transparency (XAI): While recent studies such as Erdoğmuş and Kabakuş (2023) achieve high accuracy, they require complex transformations to overcome "black box" problems. In contrast, our study offers direct explainability based on raw kinematic data, providing clinicians with a clear answer to the "why" behind a diagnosis.
3. Hardware and Resource Efficiency: Image processing-based InceptionV3 models (Erdoğmuş & Kabakuş, 2025) achieve 96.83% success but do so at the cost of millions of parameters and high GPU power requirements. The lightweight models presented in this study are capable of delivering similar accuracy on CPU-based and low-power devices, such as clinical tablets or mobile devices.

Collectively, these findings underscore that handwriting kinematics analysis stands as one of the most robust, low-cost, and objective biomarkers for the early detection of Alzheimer's Disease. This study provides a reliable foundation for the development of self-diagnosis interfaces that do not require constant specialist intervention (Kabakuş & Erdoğmuş, 2024). Future research is planned to investigate the impact of "stroke-based" microscopic analysis approaches, as proposed by Nardone et al. (2025), and to evaluate the performance of Transformer-based architectures (Kara Ardaç & Erdoğmuş, 2025) in this domain.

Ethical Considerations

Compliance with ethical guidelines

The author declares that all procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or

national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Since this study utilized a publicly available benchmark dataset (DARWIN), institutional ethics committee approval was not required.

Funding

The author declares that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The author acknowledges the use of generative AI tools for linguistic refinement and formatting assistance during the preparation of this manuscript.

REFERENCES

Bazarbekov, I., Almagambetov, B., Niyazbayev, A. & Atadjanov, B. (2025, February 20–22). Design of a smart handwriting tool for early detection of Alzheimer's Disease. International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2025), Antalya, Türkiye. https://ieeexplore.ieee.org/abstract/document/11166606

Cilia, N. D., De Stefano, C., Fontanella, F. & di Freca, A. S. (2018). An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. *Procedia Computer Science*, *141*, 466–471. https://doi.org/10.1016/j.procs.2018.10.141

Erdoğmuş, P. & Kabakuş, A. T. (2023). The promise of convolutional neural networks for the early diagnosis of the Alzheimer's disease. *Engineering Applications of Artificial Intelligence*, *123*, 106254. https://doi.org/10.1016/j.engappai.2023.106240

Erdoğmuş, P. & Kabakuş, A. T. (2025). Early diagnosis of Alzheimer's Disease using hybrid CNN-Transformer models with Grad-CAM interpretability. Gümüşhane University Journal of Science, *15*(3), 829–853. https://doi.org/10.17714/gumusfenbil.1714884

Kabakuş, A. T. & Erdoğmuş, P. (2024). Empowering self-detection: A graphical user interface powered by machine learning for early diagnosis of Alzheimer's disease. Istanbul Commerce University Journal of Science, *23*(46), 245-270. https://doi.org/10.55071/ticaretfbd.1416508

Kara Ardaç, F. B. & Erdoğmuş, P. (2025). A transformer-based method for semantic segmentation of mitosis in breast histopathology images, 9th International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Türkiye. https://ieeexplore.ieee.org/abstract/document/11222172

Lundberg, S. M. & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NIPS), *30*, 4765–4774.

Mitra, U., Chatterjee, S., Das, S., & Ahmed, K. (2024). ML-powered handwriting analysis for early detection of Alzheimer's disease. IEEE Access, *12*, 68039–68054. https://doi.org/10.1109/ACCESS.2024.3413554

Moreira, A., Ferreira, A. & Leite, N. (2025). Prediction of Alzheimer Disease on the DARWIN Dataset with Dimensionality Reduction and Explainability Techniques. Proceedings of the 1st International Conference on Explainable AI for Neural and Symbolic Methods (EXPLAINS 2024). https://doi.org/10.5220/0013017400003886

Nardone, E., De Stefano, C., Cilia, N. D., & Fontanella, F. (2025). Handwriting strokes as biomarkers for Alzheimer's disease prediction: A novel machine learning approach. *Computers in Biology and Medicine*, *190*, 110039. https://doi.org/10.1016/j.compbiomed.2024.109156

UCI Machine Learning Repository. (2024). DARWIN (Diagnosis AlzheimeR With handwritINg) htps://archive.ics.uci.edu/ml/datasets/DARWIN